5 ANGLE MODULATION AND DEMODULATION

As discussed in the previous chapter, a carrier modulation can be achieved by modulating the amplitude, frequency, and phase of a **sinusoidal carrier** of frequency f_c. In that chapter, we focused on various linear amplitude modulation systems and their demodulations. Now we discuss nonlinear frequency modulation (FM) and phase modulation (PM), often collectively known as angle modulation.

5.1 NONLINEAR MODULATION

In AM signals, the amplitude of a carrier is modulated by a signal $m(t)$, and, hence, the information content of $m(t)$ is in the amplitude variations of the carrier. As we have seen, the other two parameters of the carrier sinusoid, namely its frequency and phase, can also be varied in proportion to the message signal as frequency-modulated and phase-modulated signals, respectively. We now describe the essence of frequency modulation (FM) and phase modulation (PM).

False Start

In the 1920s, broadcasting was in its infancy. However, there was an active search for techniques to reduce noise (static). Since the noise power is proportional to the modulated signal bandwidth (sidebands), efforts were focused on finding a modulation scheme that would reduce the bandwidth. More important still, bandwidth reduction also allows more users, and there were rumors of a new method that had been discovered for eliminating sidebands (no sidebands, no bandwidth!). The idea of **frequency modulation (FM)**, where the carrier frequency would be varied in proportion to the message $m(t)$, was quite intriguing. The carrier angular frequency $\omega(t)$ would be varied with time so that $\omega(t) = \omega_c + km(t)$, where k is an arbitrary constant. If the peak amplitude of $m(t)$ is m_p, then the maximum and minimum values of the carrier frequency would be $\omega_c + km_p$ and $\omega_c - km_p$, respectively. Hence, the spectral components would remain within this band with a bandwidth $2km_p$ centered at ω_c. The understanding was that controlling the constant parameter k can control the modulated signal bandwidth. While this is true, there was also the hope that by using an arbitrarily small k, we could make the information bandwidth arbitrarily small. This possibility was seen as a passport to communication heaven. Unfortunately, experimental results showed that the underlying reasoning was seriously wrong. The FM bandwidth, as it turned out, is always greater than (at best equal to)

Figure 5.1 Concept of instantaneous frequency.

the AM bandwidth. In some cases, its bandwidth was several times that of AM. Where was the fallacy in the original reasoning? We shall soon find out.

The Concept of Instantaneous Frequency

While AM signals carry a message with their varying amplitude, FM signals can vary the instantaneous frequency in proportion to the modulating signal $m(t)$. This means that the carrier frequency is changing continuously every instant. Prima facie, this does not make much sense, since to define a frequency, we must have a sinusoidal signal at least over one cycle (or a half-cycle or a quarter-cycle) with the same frequency. This problem reminds us of our first encounter with the concept of **instantaneous velocity** in a beginning mechanics course. Until the presentation of derivatives via Leibniz and Newton, we were used to thinking of velocity as being constant over an interval, and we were incapable of even imagining that velocity could vary at each instant. We never forget, however, the wonder and amazement that was caused by the contemplation of derivative and instantaneous velocity when these concepts were first introduced. A similar experience awaits the reader with respect to **instantaneous frequency**.

Let us consider a generalized sinusoidal signal $\varphi(t)$ given by

$$\varphi(t) = A \cos \theta(t) \tag{5.1}$$

where $\theta(t)$ is the **generalized angle** and is a function of t. Figure 5.1 shows a hypothetical case of $\theta(t)$. The generalized angle for a conventional sinusoid $A \cos(\omega_c t + \theta_0)$ is a straight line $\omega_c t + \theta_0$, as shown in Fig. 5.1. A hypothetical case general angle of $\theta(t)$ happens to be tangential to the angle $(\omega_c t + \theta_0)$ at some instant t. The crucial point is that, around t, over a small interval $\Delta t \to 0$, the signal $\varphi(t) = A \cos \theta(t)$ and the sinusoid $A \cos(\omega_c t + \theta_0)$ are identical; that is,

$$\varphi(t) = A \cos(\omega_c t + \theta_0) \qquad t_1 < t < t_2$$

We are certainly justified in saying that over this small interval Δt, the angular frequency of $\varphi(t)$ is ω_c. Because $(\omega_c t + \theta_0)$ is tangential to $\theta(t)$, the angular frequency of $\varphi(t)$ is the slope of its angle $\theta(t)$ over this small interval. We can generalize this concept at **every instant** and define that the instantaneous frequency ω_i at any instant t is the slope of $\theta(t)$ at t. Thus, for $\varphi(t)$

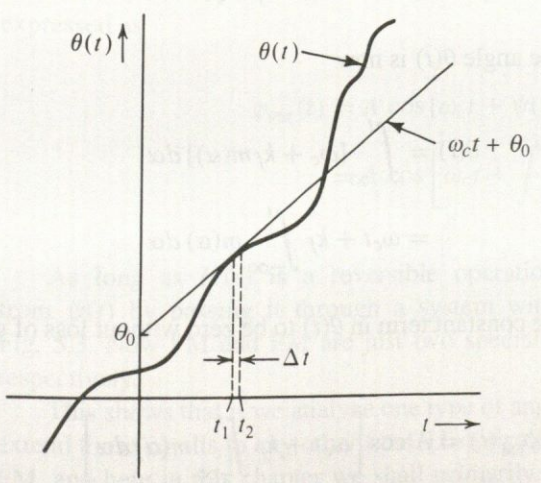

Figure 5.1 Concept of instantaneous frequency.

in Eq. (5.1), the instantaneous angular frequency and the generalized angle are related via

$$\omega_i(t) = \frac{d\theta}{dt} \tag{5.2a}$$

$$\theta(t) = \int_{-\infty}^{t} \omega_i(\alpha)\, d\alpha \tag{5.2b}$$

Figure 5.2 Phase and frequency modulation are equivalent and interchangeable.

Now we can see the possibility of transmitting the information of $m(t)$ by varying the angle θ of a carrier. Such techniques of modulation, where the angle of the carrier is varied in some manner with a modulating signal $m(t)$, are known as **angle modulation** or **exponential modulation**. Two simple possibilities are **phase modulation (PM)** and **frequency modulation (FM)**. In PM, the angle $\theta(t)$ is varied linearly with $m(t)$:

$$\theta(t) = \omega_c t + \theta_0 + k_p m(t)$$

where k_p is a constant and ω_c is the carrier frequency. Assuming $\theta_0 = 0$, without loss of generality,

$$\theta(t) = \omega_c t + k_p m(t) \tag{5.3a}$$

The resulting PM wave is

$$\varphi_{\text{PM}}(t) = A \cos\left[\omega_c t + k_p m(t)\right] \tag{5.3b}$$

The instantaneous angular frequency $\omega_i(t)$ in this case is given by

$$\omega_i(t) = \frac{d\theta}{dt} = \omega_c + k_p \dot{m}(t) \tag{5.3c}$$

Hence, in PM, the instantaneous angular frequency ω_i varies linearly with the derivative of the modulating signal. If the instantaneous frequency ω_i is varied linearly with the modulating signal, we have FM. Thus, in FM the instantaneous angular frequency ω_i is

$$\omega_i(t) = \omega_c + k_f m(t) \tag{5.4a}$$

where k_f is a constant. The angle $\theta(t)$ is now

$$\begin{aligned}\theta(t) &= \int_{-\infty}^{t} \left[\omega_c + k_f m(\alpha)\right] d\alpha \\ &= \omega_c t + k_f \int_{-\infty}^{t} m(\alpha)\, d\alpha\end{aligned}$$

Here we have assumed the constant term in $\theta(t)$ to be zero without loss of generality. The FM wave is

$$\varphi_{\text{FM}}(t) = A \cos\left[\omega_c t + k_f \int_{-\infty}^{t} m(\alpha)\, d\alpha\right] \tag{5.5}$$

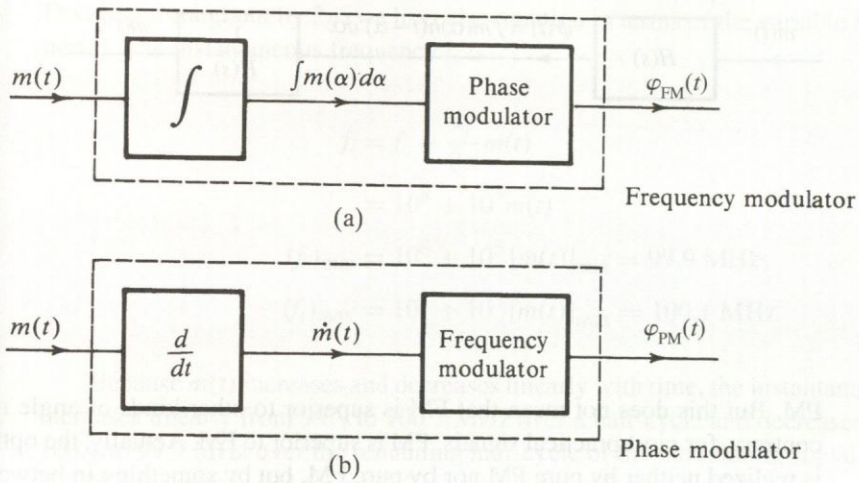

Figure 5.2 Phase and frequency modulation are equivalent and interchangeable.

Relationship between FM and PM

From Eqs. (5.3b) and (5.5), it is apparent that PM and FM not only are very similar but are inseparable. Replacing $m(t)$ in Eq. (5.3b) with $\int m(\alpha)d\alpha$ changes PM into FM. Thus, a signal that is an FM wave corresponding to $m(t)$ is also the PM wave corresponding to $\int m(\alpha)\,d\alpha$ (Fig. 5.2a). Similarly, a PM wave corresponding to $m(t)$ is the FM wave corresponding to $\dot{m}(t)$ (Fig. 5.2b). Therefore, by looking only at an angle-modulated signal $\varphi(t)$, there is no way of telling whether it is FM or PM. In fact, it is meaningless to ask an angle-modulated wave whether it is FM or PM. It is analogous to asking a married man with children whether he is a father or a son. This discussion and Fig. 5.2 also show that we need not separately discuss methods of generation and demodulation of each type of modulation.

Equations (5.3b) and (5.5) show that in both PM and FM the angle of a carrier is varied in proportion to some measure of $m(t)$. In PM, it is directly proportional to $m(t)$, whereas in FM, it is proportional to the integral of $m(t)$. As shown in Fig. 5.2b, a frequency modulator can be directly used to generate an FM signal or the message input $m(t)$ can be processed by a filter (differentiator) with transfer function $H(s) = s$ to generate PM signals. But why should we limit ourselves to these cases? We have an infinite number of possible ways of processing $m(t)$ before FM. If we restrict the choice to a linear operator, then a measure of $m(t)$ can be obtained as the output of an invertible linear (time-invariant) system with transfer function $H(s)$ or impulse response $h(t)$. The generalized angle-modulated carrier $\varphi_{EM}(t)$ can be expressed as

$$\varphi_{EM}(t) = A\cos[\omega_c t + \psi(t)] \tag{5.6a}$$

$$= A\cos\left[\omega_c t + \int_{-\infty}^{t} m(\alpha)\,h(t-\alpha)\,d\alpha\right] \tag{5.6b}$$

As long as $H(s)$ is a reversible operation (or invertible), $m(t)$ can be recovered from $\psi(t)$ by passing it through a system with transfer function $[H(s)]^{-1}$ as shown in Fig. 5.3. Now PM and FM are just two special cases with $h(t) = k_p\delta(t)$ and $h(t) = k_f u(t)$, respectively.

This shows that if we analyze one type of angle modulation (such as FM), we can readily extend those results to any other kind. Historically, the angle modulation concept began with FM, and here in this chapter we shall primarily analyze FM, with occasional discussion of

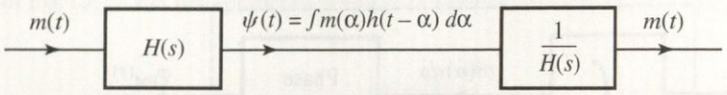

Figure 5.3 Generalized phase modulation by means of the filter $H(s)$ and recovery of the message from the modulated phase through the inverse filter $1/H(s)$.

PM. But this does not mean that FM is superior to other kinds of angle modulation. On the contrary, for most practical signals, PM is superior to FM. Actually, the optimum performance is realized neither by pure PM nor by pure FM, but by something in between.

Power of an Angle-Modulated Wave

Although the instantaneous frequency and phase of an angle-modulated wave can vary with time, the amplitude A remains constant. Hence, the power of an angle-modulated wave (PM or FM) is always $A^2/2$, regardless of the value of k_p or k_f.

Example 5.1 Sketch FM and PM waves for the modulating signal $m(t)$ shown in Fig. 5.4a. The constants k_f and k_p are $2\pi \times 10^5$ and 10π, respectively, and the carrier frequency f_c is 100 MHz.

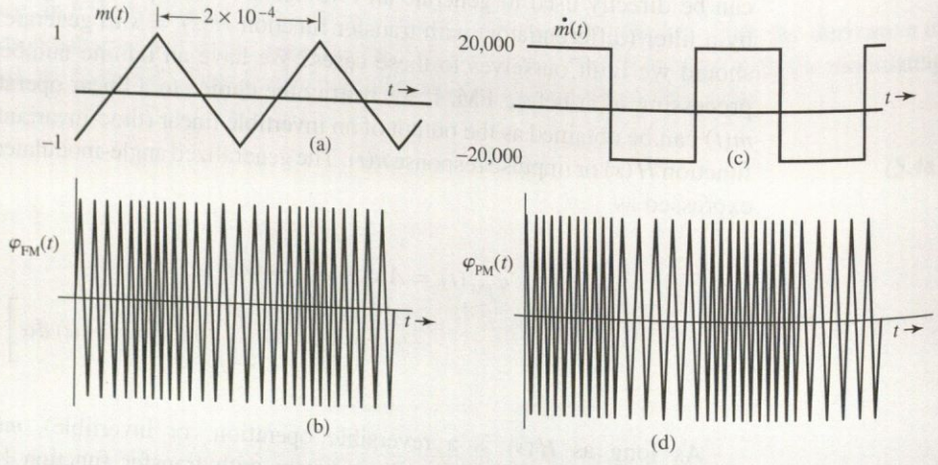

Figure 5.4 FM and PM waveforms.

For FM:

$$\omega_i = \omega_c + k_f m(t)$$

Dividing throughout by 2π, we have the equation in terms of the variable f (frequency in hertz). The instantaneous frequency f_i is

$$\begin{aligned} f_i &= f_c + \frac{k_f}{2\pi} m(t) \\ &= 10^8 + 10^5 m(t) \\ (f_i)_{\min} &= 10^8 + 10^5 [m(t)]_{\min} = 99.9 \text{ MHz} \\ (f_i)_{\max} &= 10^8 + 10^5 [m(t)]_{\max} = 100.1 \text{ MHz} \end{aligned}$$

Because $m(t)$ increases and decreases linearly with time, the instantaneous frequency increases linearly from 99.9 to 100.1 MHz over a half-cycle and decreases linearly from 100.1 to 99.9 MHz over the remaining half-cycle of the modulating signal (Fig. 5.4b).

PM for $m(t)$ is FM for $\dot{m}(t)$. This also follows from Eq. (5.3c).

For PM:

$$\begin{aligned} f_i &= f_c + \frac{k_p}{2\pi} \dot{m}(t) \\ &= 10^8 + 5\, \dot{m}(t) \\ (f_i)_{\min} &= 10^8 + 5 [\dot{m}(t)]_{\min} = 10^8 - 10^5 = 99.9 \text{ MHz} \\ (f_i)_{\max} &= 10^8 + 5 [\dot{m}(t)]_{\max} = 100.1 \text{ MHz} \end{aligned}$$

Because $\dot{m}(t)$ switches back and forth from a value of $-20{,}000$ to 20,000, the carrier frequency switches back and forth from 99.9 to 100.1 MHz every half-cycle of $\dot{m}(t)$, as shown in Fig. 5.4d.

This indirect method of sketching PM [using $\dot{m}(t)$ to frequency-modulate a carrier] works as long as $m(t)$ is a continuous signal. If $m(t)$ is discontinuous, it means that the PM signal has sudden phase changes and, hence, $\dot{m}(t)$ contains impulses. This indirect method fails at *points of the discontinuity*. In such a case, a direct approach should be used at the point of discontinuity to specify the sudden phase changes. This is demonstrated in the next example.

Example 5.2 Sketch FM and PM waves for the digital modulating signal $m(t)$ shown in Fig. 5.5a. The constants k_f and k_p are $2\pi \times 10^5$ and $\pi/2$, respectively, and $f_c = 100$ MHz.

For FM:

$$f_i = f_c + \frac{k_f}{2\pi} m(t) = 10^8 + 10^5 m(t)$$

Because $m(t)$ switches from 1 to -1 and vice versa, the FM wave frequency switches back and forth between 99.9 and 100.1 MHz, as shown in Fig. 5.5b. This scheme of carrier

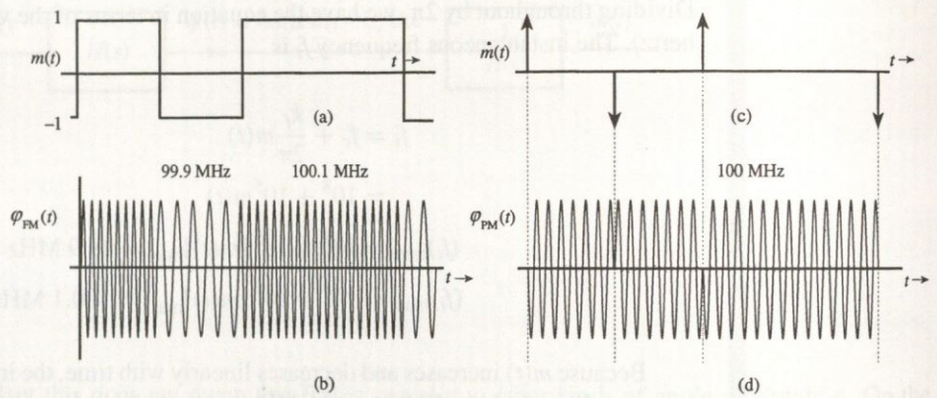

Figure 5.5 FM and PM waveforms.

frequency modulation by a digital signal (Fig. 5.5b) is called **frequency shift keying (FSK)** because information digits are transmitted by keying different frequencies (see Sec. 7.8).

For PM:

$$f_i = f_c + \frac{k_p}{2\pi}\dot{m}(t) = 10^8 + \frac{1}{4}\dot{m}(t)$$

The derivative $\dot{m}(t)$ (Fig. 5.5c) is zero except at points of discontinuity of $m(t)$ where impulses of strength ± 2 are present. This means that the frequency of the PM signal stays the same except at these isolated points of time! It is not immediately apparent how an instantaneous frequency can be changed by an infinite amount and then changed back to the original frequency in zero time. Let us consider the direct approach:

$$\begin{aligned}\varphi_{\text{PM}}(t) &= A\cos\left[\omega_c t + k_p m(t)\right] \\ &= A\cos\left[\omega_c t + \frac{\pi}{2}m(t)\right] \\ &= \begin{cases} A\sin\omega_c t & \text{when } m(t) = -1 \\ -A\sin\omega_c t & \text{when } m(t) = 1 \end{cases}\end{aligned}$$

This PM wave is shown in Fig. 5.5d. This scheme of carrier PM by a digital signal is called **phase shift keying (PSK)** because information digits are transmitted by shifting the carrier phase. Note that PSK may also be viewed as a DSB-SC modulation by $m(t)$.

The PM wave $\varphi_{\text{PM}}(t)$ in this case has phase discontinuities at instants where impulses of $\dot{m}(t)$ are located. At these instants, the carrier phase shifts by π instantaneously. A finite phase shift in zero time implies infinite instantaneous frequency at these instants. This agrees with our observation about $\dot{m}(t)$.

The amount of phase discontinuity in $\varphi_{\text{PM}}(t)$ at the instant where $m(t)$ is discontinuous is $k_p m_d$, where m_d is the amount of discontinuity in $m(t)$ at that instant. In the present example, the amplitude of $m(t)$ changes by 2 (from -1 to 1) at the discontinuity. Hence, the phase discontinuity in $\varphi_{\text{PM}}(t)$ is $k_p m_d = (\pi/2) \times 2 = \pi$ rad, which confirms our earlier result.

When $m(t)$ is a digital signal (as in Fig. 5.5a), $\varphi_{\text{PM}}(t)$ shows a phase discontinuity where $m(t)$ has a jump discontinuity. We shall now show that to avoid ambiguity in demodulation, in such a case, the phase deviation $k_p m(t)$ must be restricted to a range $(-\pi, \pi)$. For example, if k_p were $3\pi/2$ in the present example, then

$$\varphi_{\text{PM}}(t) = A \cos\left[\omega_c t + \frac{3\pi}{2} m(t)\right]$$

In this case $\varphi_{\text{PM}}(t) = A \sin \omega_c t$ when $m(t) = 1$ or $-1/3$. This will certainly cause ambiguity at the receiver when $A \sin \omega_c t$ is received. Specifically, the receiver cannot decide the exact value of $m(t)$. Such ambiguity never arises if $k_p m(t)$ is restricted to the range $(-\pi, \pi)$.

What causes this ambiguity? When $m(t)$ has jump discontinuities, the phase of $\varphi_{\text{PM}}(t)$ changes instantaneously. Because a phase $\varphi_o + 2n\pi$ is indistinguishable from the phase φ_o, ambiguities will be inherent in the demodulator unless the phase variations are limited to the range $(-\pi, \pi)$. This means k_p should be small enough to restrict the phase change $k_p m(t)$ to the range $(-\pi, \pi)$.

No such restriction on k_p is required if $m(t)$ is continuous. In this case the phase change is not instantaneous, but gradual over time, and a phase $\varphi_o + 2n\pi$ will exhibit n additional carrier cycles in the case of phase of only φ_o. We can detect the PM wave by using an FM demodulator followed by an integrator (see Prob. 5.4-1). The additional n cycles will be detected by the FM demodulator, and the subsequent integration will yield a phase $2n\pi$. Hence, the phases φ_o and $\varphi_o + 2n\pi$ can be detected without ambiguity. This conclusion can also be verified from Example 5.1, where the maximum phase change $\Delta\varphi = 10\pi$.

Because a band-limited signal cannot have jump discontinuities, we can also say that when $m(t)$ is band-limited, k_p has no restrictions.

5.2 BANDWIDTH OF ANGLE-MODULATED WAVES

Unlike AM, angle modulation is nonlinear and no properties of Fourier transform can be directly applied for its bandwidth analysis. To determine the bandwidth of an FM wave, let us define

$$a(t) = \int_{-\infty}^{t} m(\alpha)\, d\alpha \tag{5.7}$$

and define

$$\hat{\varphi}_{\text{FM}}(t) = A\, e^{j[\omega_c t + k_f a(t)]} = A e^{jk_f a(t)} e^{j\omega_c t} \tag{5.8a}$$

such that its relationship to the FM signal is

$$\varphi_{\text{FM}}(t) = \text{Re}\left[\hat{\varphi}_{\text{FM}}(t)\right] \tag{5.8b}$$

Expanding the exponential $e^{jk_f a(t)}$ of Eq. (5.8a) in power series yields

$$\hat{\varphi}_{\text{FM}}(t) = A\left[1 + jk_f a(t) - \frac{k_f^2}{2!} a^2(t) + \cdots + j^n \frac{k_f^n}{n!} a^n(t) + \cdots\right] e^{j\omega_c t} \tag{5.9a}$$

and

$$\varphi_{\text{FM}}(t) = \text{Re}\,[\hat{\varphi}_{\text{FM}}(t)]$$
$$= A\left[\cos\,\omega_c t - k_f a(t)\sin\,\omega_c t - \frac{k_f^2}{2!}a^2(t)\cos\,\omega_c t + \frac{k_f^3}{3!}a^3(t)\sin\,\omega_c t + \cdots\right] \quad (5.9b)$$

The modulated wave consists of an unmodulated carrier plus various amplitude-modulated terms, such as $a(t)\sin\,\omega_c t$, $a^2(t)\cos\,\omega_c t$, $a^3(t)\sin\,\omega_c t, \ldots$. The signal $a(t)$ is an integral of $m(t)$. If $M(f)$ is band-limited to B, $A(f)$ is also band-limited* to B. The spectrum of $a^2(t)$ is simply $A(f) * A(f)$ and is band-limited to $2B$. Similarly, the spectrum of $a^n(t)$ is band-limited to nB. Hence, the spectrum consists of an unmodulated carrier plus spectra of $a(t)$, $a^2(t), \ldots, a^n(t), \ldots$, centered at ω_c. Clearly, the modulated wave is not band-limited. It has an infinite bandwidth and is not related to the modulating-signal spectrum in any simple way, as was the case in AM.

Although the bandwidth of an FM wave is theoretically infinite, for practical signals with bounded $|a(t)|$, $|k_f a(t)|$ will remain finite. Because $n!$ increases much faster than $|k_f a(t)|^n$, we have

$$\frac{k_f^n a^n(t)}{n!} \simeq 0 \qquad \text{for large } n$$

Hence, we shall see that most of the modulated-signal power resides in a finite bandwidth. This is the principal foundation of the bandwidth analysis for angle-modulations. There are two distinct possibilities in terms of bandwidths—narrowband FM and wideband FM.

Figure 5.6 Estimation of FM wave bandwidth.

Narrowband Angle Modulation Approximation

Unlike AM, angle modulations are nonlinear. The nonlinear relationship between $a(t)$ and $\varphi(t)$ is evident from the terms involving $a^n(t)$ in Eq. (5.9b). When k_f is very small such that

$$|k_f a(t)| \ll 1$$

then all higher order terms in Eq. (5.9b) are negligible except for the first two. We then have a good approximation

$$\varphi_{\text{FM}}(t) \approx A\left[\cos\,\omega_c t - k_f a(t)\sin\,\omega_c t\right] \quad (5.10)$$

This approximation is a linear modulation that has an expression similar to that of the AM signal with message signal $a(t)$. Because the bandwidth of $a(t)$ is B Hz, the bandwidth of $\varphi_{\text{FM}}(t)$ in Eq. (5.10) is $2B$ Hz according to the frequency-shifting property due to the term $a(t)\sin\,\omega_c t$. For this reason, the FM signal for the case of $|k_f a(t)| \ll 1$ is called **narrowband FM (NBFM)**. Similarly, the **narrowband PM (NBPM)** signal is approximated by

$$\varphi_{\text{PM}}(t) \approx A\left[\cos\,\omega_c t - k_p m(t)\sin\,\omega_c t\right] \quad (5.11)$$

NBPM also has the approximate bandwidth of $2B$.

---

* This is because integration is a linear operation equivalent to passing a signal through a transfer function $1/j2\pi f$. Hence, if $M(f)$ is band-limited to B, $A(f)$ must also be band-limited to B.

A comparison of NBFM [Eq. (5.10)] with AM [Eq. (5.9a)] brings out clearly the similarities and differences between the two types of modulation. Both have the same modulated bandwidth $2B$. The sideband spectrum for FM has a phase shift of $\pi/2$ with respect to the carrier, whereas that of AM is in phase with the carrier. It must be remembered, however, that despite the apparent similarities, the AM and FM signals have very different waveforms. In an AM signal, the oscillation frequency is constant and the amplitude varies with time, whereas in an FM signal, the amplitude stays constant and the frequency varies with time.

Wideband FM (WBFM) Bandwidth Analysis: The Fallacy Exposed

Note that an FM signal is meaningful only if its frequency deviation is large enough. In other words, practical FM chooses the constant k_f large enough that the condition $|k_f a(t)| \ll 1$ is not satisfied. We call FM signals in such cases **wideband FM (WBFM)**. Thus, in analyzing the bandwidth of WBFM, we cannot ignore all the higher order terms in Eq. (5.9b). To begin, we shall take here the route of the pioneers, who by their intuitively simple reasoning came to grief in estimating the FM bandwidth. If we could discover the fallacy in their reasoning, we would have a chance of obtaining a better estimate of the (wideband) FM bandwidth.

Consider a low-pass $m(t)$ with bandwidth B Hz. This signal is well approximated by a staircase signal $\hat{m}(t)$, as shown in Fig. 5.6a. The signal $m(t)$ is now approximated by pulses of constant amplitude. For convenience, each of these pulses will be called a "cell." To ensure

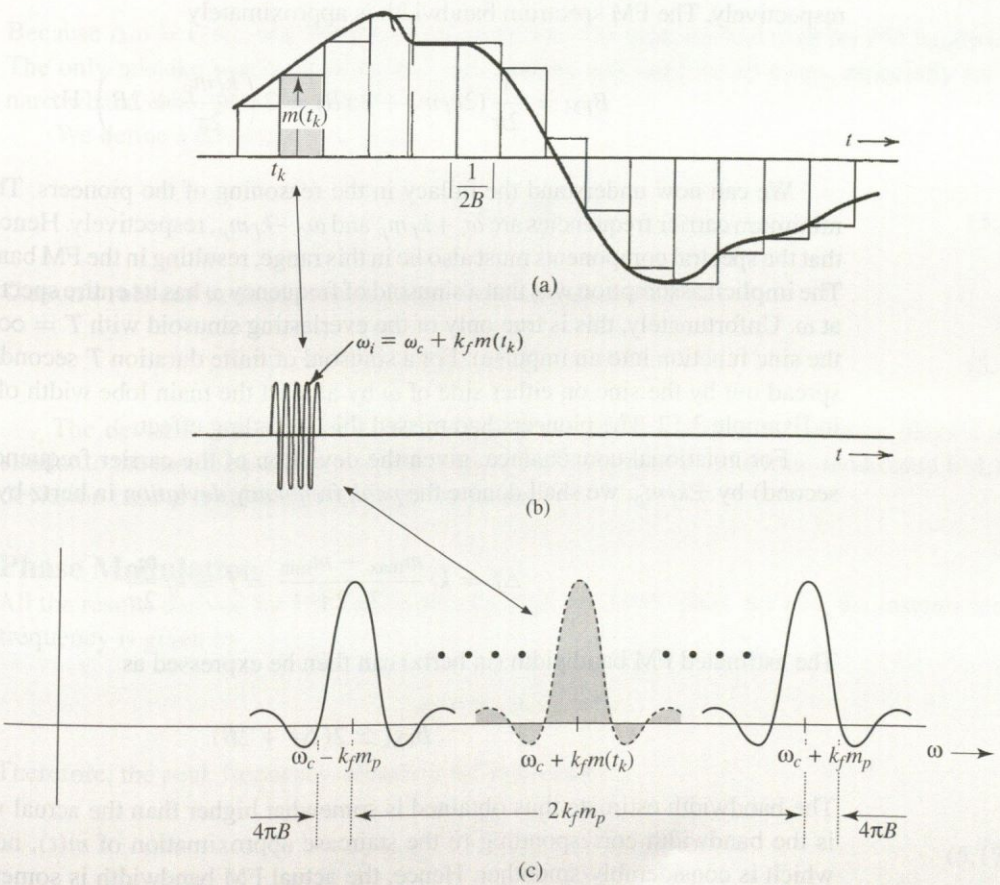

Figure 5.6 Estimation of FM wave bandwidth.

that $\hat{m}(t)$ has all the information of $m(t)$, the cell width in $\hat{m}(t)$ must be no greater than the Nyquist interval of $1/2B$ second according to the sampling theorem (Chapter 6).

It is relatively easier to analyze FM corresponding to $\hat{m}(t)$ because its constant amplitude pulses (cells) of width $T = 1/2B$ second. Consider a typical cell starting at $t = t_k$. This cell has a constant amplitude $m(t_k)$. Hence, the FM signal corresponding to this cell is a sinusoid of frequency $\omega_c + k_f m(t_k)$ and duration $T = 1/2B$, as shown in Fig. 5.6b. The FM signal for $\hat{m}(t)$ consists of a sequence of such constant frequency sinusoidal pulses of duration $T = 1/2B$ corresponding to various cells of $\hat{m}(t)$. The FM spectrum for $\hat{m}(t)$ consists of the sum of the Fourier transforms of these sinusoidal pulses corresponding to all the cells. The Fourier transform of a sinusoidal pulse in Fig. 5.6b (corresponding to the kth cell) is a sinc function shown shaded in Fig. 5.6c (see Example 3.12, Fig. 3.22d with $T = 1/2B$).

$$\text{rect}(2Bt)\cos[\omega_c t + k_f m(t_k)t] \Longleftrightarrow \frac{1}{2}\text{sinc}\left[\frac{\omega + \omega_c + k_f m(t_k)}{4B}\right] + \frac{1}{2}\text{sinc}\left[\frac{\omega - \omega_c - k_f m(t_k)}{4B}\right]$$

Note that the spectrum of this pulse is spread out on either side of its center frequency $\omega_c + k_f m(t_k)$ by $4\pi B$ as the main lobe of the sinc function. Figure 5.6c shows the spectra of sinusoidal pulses corresponding to various cells. The minimum and the maximum amplitudes of the cells are $-m_p$ and m_p, respectively. Hence, the minimum and maximum *center* frequencies of the short sinusoidal pulses corresponding to the FM signal for all the cells are $\omega_c - k_f m_p$ and $\omega_c + k_f m_p$, respectively. Consider the sinc main lobe of these frequency responses as significant contribution to the FM bandwidth, as shown in Fig. 5.6c. Hence, the maximum and the minimum significant frequencies in this spectrum are $\omega_c + k_f m_p + 4\pi B$ and $\omega_c - k_f m_p - 4\pi B$, respectively. The FM spectrum bandwidth is approximately

$$B_{\text{FM}} = \frac{1}{2\pi}(2k_f m_p + 8\pi B) = 2\left(\frac{k_f m_p}{2\pi} + 2B\right) \text{ Hz}$$

We can now understand the fallacy in the reasoning of the pioneers. The maximum and minimum carrier frequencies are $\omega_c + k_f m_p$ and $\omega_c - k_f m_p$, respectively. Hence, it was reasoned that the spectral components must also lie in this range, resulting in the FM bandwidth of $2k_f m_p$. The implicit assumption was that a sinusoid of frequency ω has its entire spectrum concentrated at ω. Unfortunately, this is true only of the everlasting sinusoid with $T = \infty$ (because it turns the sinc function into an impulse). For a sinusoid of finite duration T seconds, the spectrum is spread out by the sinc on either side of ω by at least the main lobe width of $2\pi/T$, as shown in Example 3.12. The pioneers had missed this spreading effect.

For notational convenience, given the deviation of the carrier frequency (in radians per second) by $\pm k_f m_p$, we shall denote the *peak frequency deviation* in hertz by Δf. Thus,

$$\Delta f = k_f \frac{m_{\max} - m_{\min}}{2 \cdot 2\pi} = f = k_f \frac{m_p}{2\pi}$$

The estimated FM bandwidth (in hertz) can then be expressed as

$$B_{\text{FM}} \simeq 2(\Delta f + 2B) \tag{5.12}$$

The bandwidth estimate thus obtained is somewhat higher than the actual value because this is the bandwidth corresponding to the staircase approximation of $m(t)$, not the actual $m(t)$, which is considerably smoother. Hence, the actual FM bandwidth is somewhat smaller than

this value. Based on Fig. 5.6c, it is clear that a better FM bandwidth approximation is between

$$[2\Delta f,\ 2\Delta f + 4B]$$

Therefore, we should readjust our bandwidth estimation. To make this midcourse correction, we observe that for the case of NBFM, k_f is very small. Hence, given a fixed m_p, Δf is very small (in comparison to B) for NBFM. In this case, we can ignore the small Δf term in Eq. (5.12) with the result

$$B_{\text{FM}} \approx 4B$$

But we showed earlier that for narrowband, the FM bandwidth is approximately $2B$ Hz. This indicates that a better bandwidth estimate is

$$B_{\text{FM}} = 2(\Delta f + B) = 2\left(\frac{k_f m_p}{2\pi} + B\right) \tag{5.13}$$

This is precisely the result obtained by Carson,[1] who investigated this problem rigorously for tone modulation [sinusoidal $m(t)$]. This formula goes under the name **Carson's rule** in the literature. Observe that for a truly wideband case, where $\Delta f \gg B$, Eq. (5.13) can be approximated as

$$B_{\text{FM}} \approx 2\Delta f \qquad \Delta f \gg B \tag{5.14}$$

Because $\Delta\omega = k_f m_p$, this formula is precisely what the pioneers had used for FM bandwidth. The only mistake was in thinking that this formula will hold for all cases, especially for the narrowband case, where $\Delta f \ll B$.

We define a deviation ratio β as

$$\beta = \frac{\Delta f}{B} \tag{5.15}$$

Carson's rule can be expressed in terms of the deviation ratio as

$$B_{\text{FM}} = 2B(\beta + 1) \tag{5.16}$$

The deviation ratio controls the amount of modulation and, consequently, plays a role similar to the modulation index in AM. Indeed, for the special case of tone-modulated FM, the deviation ratio β is called the **modulation index.**

Phase Modulation

All the results derived for FM can be directly applied to PM. Thus, for PM, the instantaneous frequency is given by

$$\omega_i = \omega_c + k_p \dot{m}(t)$$

Therefore, the peak frequency deviation Δf is given by

$$\Delta f = k_p \frac{[\dot{m}(t)]_{\max} - [\dot{m}(t)_{\min}]}{2 \cdot 2\pi} \tag{5.17a}$$

If we assume that

$$\dot{m}_p = [\dot{m}(t)]_{\max} = -|\dot{m}(t)_{\min}| \tag{5.17b}$$

then

$$\Delta f = k_p \frac{\dot{m}_p}{2\pi} \tag{5.17c}$$

Therefore,*

$$B_{\text{PM}} = 2(\Delta f + B) \tag{5.18a}$$

$$= 2\left(\frac{k_p \dot{m}_p}{2\pi} + B\right) \tag{5.18b}$$

One very interesting aspect of FM is that $\Delta\omega = k_f m_p$ depends only on the peak value of $m(t)$. It is independent of the spectrum of $m(t)$. On the other hand, in PM, $\Delta\omega = k_p \dot{m}_p$ depends on the peak value of $\dot{m}(t)$. But $\dot{m}(t)$ depends strongly on the spectral composition of $m(t)$. The presence of higher frequency components in $m(t)$ implies rapid time variations, resulting in a higher value of $\dot{m}_p$. Conversely, predominance of lower frequency components will result in a lower value of $\dot{m}_p$. Hence, whereas the FM signal bandwidth [Eq. (5.13)] is practically independent of the spectral shape of $m(t)$, the PM signal bandwidth [Eq. (5.18)] is strongly affected by the spectral shape of $m(t)$. For $m(t)$ with a spectrum concentrated at lower frequencies, B_{PM} will be smaller than when the spectrum of $m(t)$ is concentrated at higher frequencies.

Spectral Analysis of Tone Frequency Modulation

For an FM carrier with a generic message signal $m(t)$, the spectral analysis requires the use of staircase signal approximation. Tone modulation is a special case for which a precise spectral analysis is possible: that is, when $m(t)$ is a sinusoid. We use this special case to verify the FM bandwidth approximation. Let

$$m(t) = \alpha \cos \omega_m t$$

From Eq. (5.7), with the assumption that initially $a(-\infty) = 0$, we have

$$a(t) = \frac{\alpha}{\omega_m} \sin \omega_m t$$

Thus, from Eq. (5.8a), we have

$$\hat{\varphi}_{\text{FM}}(t) = Ae^{(j\omega_c t + k_f \alpha/\omega_m \sin \omega_m t)}$$

Moreover

$$\Delta\omega = k_f m_p = \alpha k_f$$

* Equation (5.17a) can be applied only if $m(t)$ is a continuous function of time. If $m(t)$ has jump discontinuities, its derivative does not exist. In such a case, we should use the direct approach (discussed in Example 5.2) to find $\varphi_{\text{PM}}(t)$ and then determine $\Delta\omega$ from $\varphi_{\text{PM}}(t)$.

and the bandwidth of $m(t)$ is $2\pi B = \omega_m$ rad/s. The deviation ratio (or in this case, the modulation index) is

$$\beta = \frac{\Delta f}{B} = \frac{\Delta\omega}{2\pi B} = \frac{\alpha k_f}{\omega_m}$$

Hence,

$$\begin{aligned}\hat{\varphi}_{\text{FM}}(t) &= Ae^{j(\omega_c t + \beta \sin \omega_m t)} \\ &= Ae^{j\omega_c t}(e^{j\beta \sin \omega_m t})\end{aligned} \tag{5.19}$$

Note that $e^{j\beta \sin \omega_m t}$ is a periodic signal with period $2\pi/\omega_m$ and can be expanded by the exponential Fourier series, as usual,

$$e^{j\beta \sin \omega_m t} = \sum_{n=-\infty}^{\infty} D_n e^{jn\omega_m t}$$

where

$$D_n = \frac{\omega_m}{2\pi}\int_{-\pi/\omega_m}^{\pi/\omega_m} e^{j\beta \sin \omega_m t} e^{-jn\omega_m t}\, dt = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{j(\beta \sin x - nx)}\, dx$$

The integral on the right-hand side cannot be evaluated in a closed form but must be integrated by expanding the integrand in infinite series. This integral has been extensively tabulated and is denoted by $J_n(\beta)$, the Bessel function of the first kind and the nth order. These functions are plotted in Fig. 5.7a as a function of n for various values of β. Thus,

$$e^{j\beta \sin \omega_m t} = \sum_{n=-\infty}^{\infty} J_n(\beta) e^{jn\omega_m t} \tag{5.20}$$

Substituting Eq. (5.20) into Eq. (5.19), we get

$$\hat{\varphi}_{\text{FM}}(t) = A\sum_{n=-\infty}^{\infty} J_n(\beta) e^{j(\omega_c t + n\omega_m t)}$$

and

$$\varphi_{\text{FM}}(t) = A\sum_{n=-\infty}^{\infty} J_n(\beta)\cos(\omega_c + n\omega_m)t$$

The tone-modulated FM signal has a carrier component and an infinite number of sidebands of frequencies $\omega_c \pm \omega_m$, $\omega_c \pm 2\omega_m, \ldots, \omega_c \pm n\omega_m, \ldots$, as shown in Fig. 5.7b. This is in stark contrast to the DSB-SC spectrum of only one sideband on either side of the carrier frequency. The strength of the nth sideband at $\omega = \omega_c + n\omega_m$ is* $J_n(\beta)$. From the plots of $J_n(\beta)$ in Fig. 5.7a, it can be seen that for a given β, $J_n(\beta)$ decreases with n, and there are only a finite number

* Also $J_{-n}(\beta) = (-1)^n J_n(\beta)$. Hence, the magnitude of the LSB at $\omega = \omega_c - n\omega_m$ is the same as that of the USB at $\omega = \omega_c + n\omega_m$.

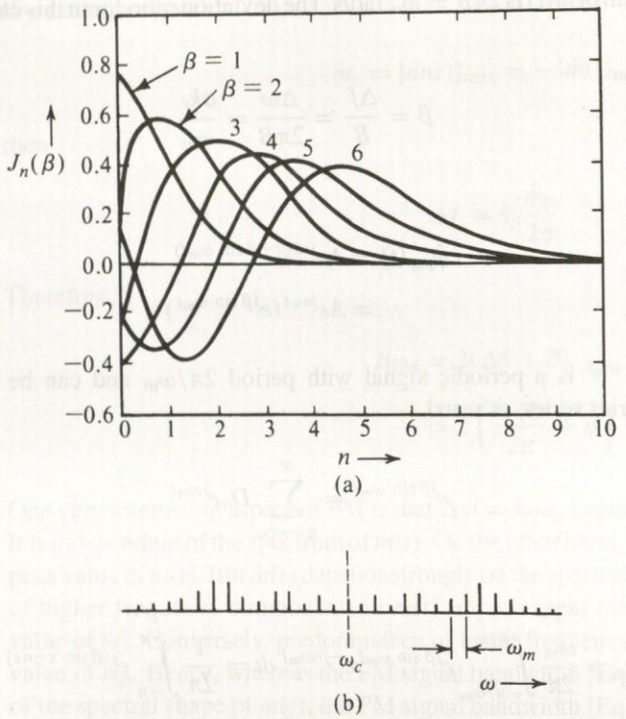

Figure 5.7 (a) Variations of $J_n(\beta)$ as a function of n for various values of β. (b) Tone-modulated FM wave spectrum.

of significant sideband spectral lines. It can be seen from Fig. 5.7a that $J_n(\beta)$ is negligible for $n > \beta + 1$. Hence, the number of significant sideband impulses is $\beta + 1$. The bandwidth of the FM carrier is given by

$$\begin{aligned} B_{\text{FM}} &= 2(\beta + 1)f_m \\ &= 2(\Delta f + B) \end{aligned}$$

which corroborates our previous result [Eqs. (5.13)]. When $\beta \ll 1$ (NBFM), there is only one significant sideband and the bandwidth $B_{\text{FM}} = 2f_m = 2B$. It is important to note that this tone modulation case analysis is a verification, not a proof, of Carson's formula.

In the literature, tone modulation in FM is often discussed in great detail. Since, however, angle modulation is a nonlinear modulation, the results derived for tone modulation may have little connection to practical situations. Indeed, these results are meaningless at best and misleading at worst when generalized to practical signals.* As authors and instructors, we feel that too much emphasis on tone modulation can be misleading. For this reason we have omitted further such discussion here.

The method for finding the spectrum of a tone-modulated FM wave can be used for finding the spectrum of an FM wave when $m(t)$ is a general periodic signal. In this case,

$$\hat{\varphi}_{\text{FM}}(t) = Ae^{j\omega_c t}[e^{jk_f a(t)}]$$

* For instance, based on tone modulation analysis, it is often stated that FM is superior to PM by a factor of 3 in terms of the output SNR. This is in fact untrue for most of the signals encountered in practices.

Because $a(t)$ is a periodic signal, $e^{jk_f a(t)}$ is also a periodic signal, which can be expressed as an exponential Fourier series in the preceding expression. After this, it is relatively straightforward to write $\varphi_{\text{FM}}(t)$ in terms of the carrier and the sidebands.

Example 5.3

(a) Estimate B_{FM} and B_{PM} for the modulating signal $m(t)$ in Fig. 5.4a for $k_f = 2\pi \times 10^5$ and $k_p = 5\pi$. Assume the essential bandwidth of the periodic $m(t)$ as the frequency of its third harmonic.

(b) Repeat the problem if the amplitude of $m(t)$ is doubled [if $m(t)$ is multiplied by 2].

(a) The peak amplitude of $m(t)$ is unity. Hence, $m_p = 1$. We now determine the essential bandwidth B of $m(t)$. It is left as an exercise for the reader to show that the Fourier series for this periodic signal is given by

$$m(t) = \sum_n C_n \cos n\omega_0 t \qquad \omega_0 = \frac{2\pi}{2 \times 10^{-4}} = 10^4\pi$$

where

$$C_n = \begin{cases} \dfrac{8}{\pi^2 n^2} & n \text{ odd} \\ 0 & n \text{ even} \end{cases}$$

It can be seen that the harmonic amplitudes decrease rapidly with n. The third harmonic is only 11% of the fundamental, and the fifth harmonic is only 4% of the fundamental. This means the third and fifth harmonic powers are 1.21 and 0.16%, respectively, of the fundamental component power. Hence, we are justified in assuming the essential bandwidth of $m(t)$ as the frequency of its third harmonic, that is,

$$B = 3 \times \frac{10^4}{2} = 15\,\text{kHz}$$

For FM:

$$\Delta f = \frac{1}{2\pi} k_f m_p = \frac{1}{2\pi}(2\pi \times 10^5)(1) = 100$$

and

$$B_{\text{FM}} = 2(\Delta f + B) = 230\,\text{kHz}$$

Alternatively, the deviation ratio β is given by

$$\beta = \frac{\Delta f}{B} = \frac{100}{15}$$

and

$$B_{\text{FM}} = 2B(\beta + 1) = 30\left(\frac{100}{15} + 1\right) = 230\,\text{kHz}$$

For PM: The peak amplitude of $\dot{m}(t)$ is 20,000 and

$$\Delta f = \frac{1}{2\pi} k_p \dot{m}_p = 50 \text{ kHz}$$

Hence,

$$B_{\text{PM}} = 2(\Delta f + B) = 130 \text{ kHz}$$

Alternately, the deviation ratio β is given by

$$\beta = \frac{\Delta f}{B} = \frac{50}{15}$$

and

$$B_{\text{PM}} = 2B(\beta + 1) = 30\left(\frac{50}{15} + 1\right) = 130 \text{ kHz}.$$

(b) Doubling $m(t)$ doubles its peak value. Hence, $m_p = 2$. But its bandwidth is unchanged so that $B = 15$ kHz.

For FM:

$$\Delta f = \frac{1}{2\pi} k_f m_p = \frac{1}{2\pi}(2\pi \times 10^5)(2) = 200 \text{ kHz}$$

and

$$B_{\text{FM}} = 2(\Delta f + B) = 430 \text{ kHz}$$

Alternately, the deviation ratio β is given by

$$\beta = \frac{\Delta f}{B} = \frac{200}{15}$$

and

$$B_{\text{FM}} = 2B(\beta + 1) = 30\left(\frac{200}{15} + 1\right) = 430 \text{ kHz}$$

For PM: Doubling $m(t)$ doubles its derivative so that now $\dot{m}_p = 40,000$, and

$$\Delta f = \frac{1}{2\pi} k_p \dot{m}_p = 100 \text{ kHz}$$

and

$$B_{\text{PM}} = 2(\Delta f + B) = 230 \text{ kHz}$$

Alternately, the deviation ratio β is given by

$$\beta = \frac{\Delta f}{B} = \frac{100}{15}$$

and

$$B_{\text{PM}} = 2B(\beta + 1) = 30\left(\frac{100}{15} + 1\right) = 230\,\text{kHz}$$

Observe that doubling the signal amplitude [doubling $m(t)$] roughly doubles frequency deviation Δf of both FM and PM waveforms.

Example 5.4 Repeat Example 5.1 if $m(t)$ is time-expanded by a factor of 2: that is, if the period of $m(t)$ is 4×10^{-4}.

Recall that time expansion of a signal by a factor of 2 reduces the signal spectral width (bandwidth) by a factor of 2. We can verify this by observing that the fundamental frequency is now 2.5 kHz, and its third harmonic is 7.5 kHz. Hence, $B = 7.5$ kHz, which is half the previous bandwidth. Moreover, time expansion does not affect the peak amplitude and thus $m_p = 1$. However, $\dot{m}_p$ is halved, that is, $\dot{m}_p = 10,000$.

For FM:

$$\Delta f = \frac{1}{2\pi} k_f m_p = 100\,\text{kHz}$$

$$B_{\text{FM}} = 2(\Delta f + B) = 2(100 + 7.5) = 215\,\text{kHz}$$

For PM:

$$\Delta f = \frac{1}{2\pi} k_p \dot{m}_p = 25\,\text{kHz}$$

$$B_{\text{PM}} = 2(\Delta f + B) = 65\,\text{kHz}$$

Note that time expansion of $m(t)$ has very little effect on the FM bandwidth, but it halves the PM bandwidth. This verifies our observation that the PM spectrum is strongly dependent on the spectrum of $m(t)$.

Example 5.5 An angle-modulated signal with carrier frequency $\omega_c = 2\pi \times 10^5$ is described by the equation

$$\varphi_{\text{EM}}(t) = 10\cos(\omega_c t + 5\,\sin\,3000t + 10\,\sin\,2000\pi t)$$

(a) Find the power of the modulated signal.
(b) Find the frequency deviation Δf.
(c) Find the deviation ratio β.

(d) Find the phase deviation $\Delta\phi$.
(e) Estimate the bandwidth of $\varphi_{\text{EM}}(t)$.

The signal bandwidth is the highest frequency in $m(t)$ (or its derivative). In this case $B = 2000\pi/2\pi = 1000$ Hz.

(a) The carrier amplitude is 10, and the power is

$$P = \frac{10^2}{2} = 50$$

(b) To find the frequency deviation Δf, we find the instantaneous frequency ω_i, given by

$$\omega_i = \frac{d}{dt}\theta(t) = \omega_c + 15{,}000 \cos 3000t + 20{,}000\pi \cos 2000\pi t$$

The carrier deviation is $15{,}000 \cos 3000t + 20{,}000\pi \cos 2000\pi t$. The two sinusoids will add in phase at some point, and the maximum value of this expression is $15{,}000+20{,}000\pi$. This is the maximum carrier deviation $\Delta\omega$. Hence,

$$\Delta f = \frac{\Delta\omega}{2\pi} = 12{,}387.32 \text{ Hz}$$

(c) $\beta = \dfrac{\Delta f}{B} = \dfrac{12{,}387.32}{1000} = 12.387$

(d) The angle $\theta_c(t) = \omega t + (5 \sin 3000t + 10 \sin 2000\pi t)$. The phase deviation is the maximum value of the angle inside the parentheses, and is given by $\Delta\phi = 15$ rad.

(e) $B_{\text{EM}} = 2(\Delta f + B) = 26{,}774.65$ Hz

Observe the generality of this method of estimating the bandwidth of an angle-modulated waveform. We need not know whether it is FM, PM, or some other kind of angle modulation. It is applicable to any angle-modulated signal.

A Historical Note: Edwin H. Armstrong (1890–1954)

Today, nobody doubts that FM has a key place in broadcasting and communication. As recently as the 1960s, however, the FM broadcasting seemed doomed because it was so uneconomical in bandwidth usage.

The history of FM is full of strange ironies. The impetus behind the development of FM was the desire to reduce signal transmission bandwidth. Superficial reasoning showed that it was feasible to reduce the transmission bandwidth by using FM. But the experimental results showed otherwise. The transmission bandwidth of FM was actually larger than that of AM. Careful mathematical analysis by Carson showed that FM indeed required a larger bandwidth than AM. Unfortunately, Carson did not recognize the compensating advantage of FM in its ability to suppress noise. Without much basis, he concluded that FM introduces inherent distortion and has no compensating advantages whatsoever.[1] In a later paper, he continues

Edwin H. Armstrong. (Reproduced with permission from Armstrong Family Archives.)

Edwin H. Armstrong. (Reproduced with permission from Armstrong Family Archives.)

"In fact, as more and more schemes are analyzed and tested, and as the essential nature of the problem is more clearly perceivable, we are unavoidably forced to the conclusion that static (noise), like the poor, will always be with us."[2] The opinion of one of the most able mathematicians of the day in the communication field, thus, set back the development of FM by more than a decade. The *noise-suppressing advantage* of FM was later proved by Major Edwin H. Armstrong,[3] a brilliant engineer whose contributions to the field of radio systems are comparable to those of Hertz and Marconi. It was largely the work of Armstrong that was responsible for rekindling the interest in FM.

Although Armstrong did not invent the concept, he has been considered the father of modern FM. Born on December 18, 1890, in New York City, Edwin H. Armstrong is widely regarded as one of the foremost contributors to radio electronics of the twentieth century. Armstrong was credited with the invention of the *regenerative circuit* (U.S. Patent 1,113,149 issued in 1912, while he was a junior at Columbia University), the *superheterodyne circuit* (U.S. Patent 1,342,885 issued in 1918, while serving in the U.S. Army stationed in Paris, during World War I), the *super-regenerative circuit* (U.S. Patent 1,424,065, issued in 1922), and the complete FM radio broadcasting system (U.S. Patent 1,941,066, 1933). All are breakthrough contributions to the radio field. *Fortune* magazine in 1939 declared: Wideband frequency modulation is the fourth, and perhaps the greatest, in a line of Armstrong inventions that have made most of modern broadcasting what it is. Major Armstrong is the acknowledged inventor of the regenerative 'feedback' circuit, which brought radio art out of the crystal-detector headphone stage and made the amplification of broadcasting possible; the superheterodyne circuit, which is the basis of practically all modern radio; and the super-regenerative circuit now in wide use in . . . shortwave systems.[4]

Armstrong was the last of the breed of the lone attic inventors. After receiving his FM patents in 1933, he gave his now famous paper (which later appeared in print as in the proceedings of the IRE[3]), accompanied by the first public demonstration of FM broadcasting on November 5, 1935, at the New York section meeting of the Institute of Radio Engineers (IRE, a predecessor of the IEEE). His success in dramatically reducing static noise using FM was not fully embraced by the broadcast establishment, which perceived FM as a threat to its vast commercial investment in AM radio. To establish FM broadcasting, Armstrong fought a

long and costly battle with the radio broadcast establishment, which, abetted by the Federal Communications Commission (FCC), fought tooth and nail to resist FM. Still, by December 1941, 67 commercial FM stations had been authorized with as many as half a million receivers in use and 43 applications were pending. In fact, the Radio Technical Planning Board (RTPB) made its final recommendation during the September 1944 FCC hearing that FM be given 75 channels in the band from 41 to 56 MHz.

Despite the recommendation of the RTPB, which was supposed to be the best advice available from the radio engineering community, strong lobbying for the FCC to shift the FM band persisted, mainly by those who propagated the concern that strong radio interferences in the 40 MHz band might be possible as a result ionospheric reflection. Then in June 1945, the FCC, on the basis of erroneous testimony of a technical expert, abruptly shifted the allocated bandwidth of FM from the 42- to 50-MHz range to the 88- to 108-MHz. This dealt a crippling blow to FM by making obsolete more than half a million receivers and equipment (transmitters, antennas, etc.) that had been built and sold by the FM industry to 50 FM stations since 1941 for the 42 to 50 MHz band. Armstrong fought the decision, and later succeeded in getting the technical expert to admit his error. In spite of all this, the FCC allocations remained unchanged. Armstrong spent the sizable fortune he had made from his inventions in legal struggles. The broadcast giants, which had so strongly resisted FM, turned around and used his inventions without paying him royalties. Armstrong spent much of his time in court in some of the longest, most notable, and acrimonious patent suits of the era.[5] In the end, with his funds depleted, his energy drained, and his family life shattered, a despondent Armstrong committed suicide: (in 1954) he walked out of a window of his thirteenth floor apartment in New York's River House.

Armstrong's widow continued the legal battles and won. By the 1960s, FM was clearly established as the superior radio system,[6] and Edwin H. Armstrong was fully recognized as the inventor of frequency modulation. In 1955 the ITU added him to its roster of great inventors. In 1980 Edwin H. Armstrong was inducted into the U.S. National Inventors Hall of Fame, and his picture was put on a U.S. postage stamp in 1983.[7]

Figure 5.8 (a) Narrowband PM generator. (b) Narrowband FM signal generator.

Figure 5.9 (a) Hard limiter and bandpass filter used to remove amplitude variations in FM wave. (b) Hard limiter input-output characteristic. (c) Hard limiter input and the corresponding output. (d) Hard limiter output as a function of θ.

5.3 GENERATING FM WAVES

Basically, there are two ways of generating FM waves: **indirect** and **direct**. We first describe the narrowband FM generator that is utilized in the **indirect FM generation** of wideband angle modulation signals.

NBFM Generation

For NBFM and NBPM signals, we have shown earlier that because $|k_f a(t)| \ll 1$ and $|k_p m(t)| \ll 1$, respectively, the modulated signals can be approximated by

$$\varphi_{\text{NBFM}}(t) \simeq A[\cos \omega_c t - k_f a(t) \sin \omega_c t] \tag{5.21a}$$

$$\varphi_{\text{NBPM}}(t) \simeq A[\cos \omega_c t - k_p m(t) \sin \omega_c t] \tag{5.21b}$$

Both approximations are linear and are similar to the expression of the AM wave. In fact, Eqs. (5.21) suggest a possible method of generating narrowband FM and PM signals by using DSB-SC modulators. The block diagram representation of such systems appears in Fig. 5.8.

It is important to point out that the NBFM generated by Fig. 5.8b has some distortion because of the approximation in Eq. (5.10). The output of this NBFM modulator also has some amplitude variations. A nonlinear device designed to limit the amplitude of a bandpass signal can remove most of this distortion.

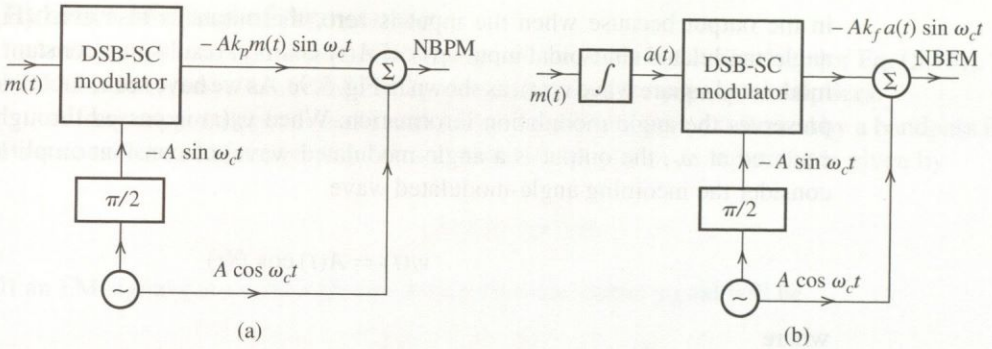

Figure 5.8
(a) Narrowband PM generator. (b) Narrowband FM signal generator.

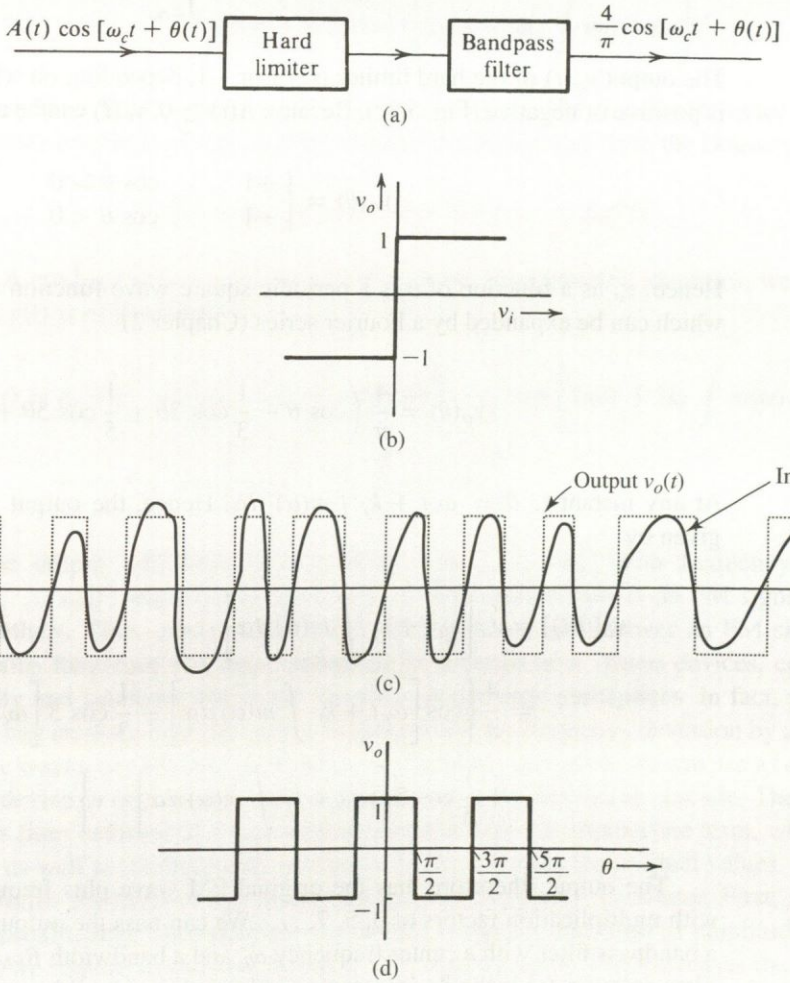

Figure 5.9
(a) Hard limiter and bandpass filter used to remove amplitude variations in FM wave. (b) Hard limiter input-output characteristic. (c) Hard limiter input and the corresponding output. (d) Hard limiter output as a function of θ.

Bandpass Limiter

The amplitude variations of an angle-modulated carrier can be eliminated by what is known as a **bandpass limiter**, which consists of a hard limiter followed by a bandpass filter (Fig. 5.9a). The input-output characteristic of a hard limiter is shown in Fig. 5.9b. Observe that the bandpass limiter output to a sinusoid will be a square wave of unit amplitude regardless of the incoming sinusoidal amplitude. Moreover, the zero crossings of the incoming sinusoid are preserved

in the output because when the input is zero, the output is also zero (Fig. 5.9b). Thus an angle-modulated sinusoidal input $v_i(t) = A(t) \cos \theta(t)$ results in a constant amplitude, angle-modulated square wave $v_o(t)$, as shown in Fig. 5.9c. As we have seen, such a nonlinear operation preserves the angle modulation information. When $v_o(t)$ is passed through a bandpass filter centered at ω_c, the output is a angle-modulated wave, of constant amplitude. To show this, consider the incoming angle-modulated wave

$$v_i(t) = A(t) \cos \theta(t)$$

where

$$\theta(t) = \omega_c t + k_f \int_{-\infty}^{t} m(\alpha)\, d\alpha$$

The output $v_o(t)$ of the hard limiter is $+1$ or -1, depending on whether $v_i(t) = A(t) \cos \theta(t)$ is positive or negative (Fig. 5.9c). Because $A(t) \geq 0$, $v_o(t)$ can be expressed as a function of θ:

$$v_o(\theta) = \begin{cases} +1 & \cos \theta > 0 \\ -1 & \cos \theta < 0 \end{cases}$$

Hence, v_o as a function of θ is a periodic square wave function with period 2π (Fig. 5.9d), which can be expanded by a Fourier series (Chapter 2)

$$v_o(\theta) = \frac{4}{\pi}\left(\cos \theta - \frac{1}{3} \cos 3\theta + \frac{1}{5} \cos 5\theta + \cdots\right)$$

At any instant t, $\theta = \omega_c t + k_f \int m(\alpha)\, d\alpha$. Hence, the output v_o as a function of time is given by

$$\begin{aligned} v_o[\theta(t)] &= v_o\left[\omega_c t + k_f \int m(\alpha)\, d\alpha\right] \\ &= \frac{4}{\pi}\left\{\cos\left[\omega_c t + k_f \int m(\alpha)\, d\alpha\right] - \frac{1}{3} \cos 3\left[\omega_c t + k_f \int m(\alpha)\, d\alpha\right]\right. \\ &\quad \left. + \frac{1}{5} \cos 5\left[\omega_c t + k_f \int m(\alpha)\, d\alpha\right] \cdots\right\} \end{aligned}$$

The output, therefore, has the original FM wave plus frequency-multiplied FM waves with multiplication factors of 3, 5, 7, ... We can pass the output of the hard limiter through a bandpass filter with a center frequency ω_c and a bandwidth B_{FM}, as shown in Fig. 5.9a. The filter output $e_o(t)$ is the desired angle-modulated carrier with a constant amplitude,

$$e_o(t) = \frac{4}{\pi} \cos\left[\omega_c(t) + k_f \int m(\alpha) d\alpha\right]$$

Although we derived these results for FM, this applies to PM (angle modulation in general) as well. The bandpass filter not only maintains the constant amplitude of the angle-modulated carrier but also partially suppresses the channel noise when the noise is small.[8]

Indirect Method of Armstrong

In Armstrong's indirect method, NBFM is generated as shown in Fig. 5.8b [or Eq. (5.10)]. The NBFM is then converted to WBFM by using additional **frequency multipliers**.

A frequency multiplier can be realized by a nonlinear device followed by a bandpass filer. First consider a nonlinear device whose output signal $y(t)$ to an input $x(t)$ is given by

$$y(t) = a_2 x^2(t)$$

If an FM signal passes through this device, then the output signal will be

$$\begin{aligned} y(t) &= a_2 \cos^2\left[\omega_c t + k_f \int m(\alpha)\, d\alpha\right] \\ &= 0.5a_2 + 0.5a_2 \cos\left[2\omega_c t + 2k_f \int m(\alpha)\, d\alpha\right] \end{aligned} \tag{5.22}$$

Thus, a bandpass filter centered at $2\omega_c$ would recover an FM signal with twice the original instantaneous frequency. To generalize, a nonlinear device may have the characteristic of

$$y(t) = a_0 + a_1 x(t) + a_2 x^2(t) + \cdots + a_n x^n(t) \tag{5.23}$$

If $x(t) = A \cos\left[\omega_c t + k_f \int m(\alpha)\, d\alpha\right]$, then by using trigonometric identities, we can readily show that $y(t)$ is of the form

$$\begin{aligned} y(t) = c_o + c_1 \cos\left[\omega_c t + k_f \int m(\alpha)\, d\alpha\right] + c_2 \cos\left[2\omega_c t + 2k_f \int m(\alpha)\, d\alpha\right] \\ + \cdots + c_n \cos\left[n\omega_c t + nk_f \int m(\alpha)\, d\alpha\right] \end{aligned} \tag{5.24}$$

Hence, the output will have spectra at ω_c, $2\omega_c, \ldots,$ $n\omega_c$, with frequency deviations Δf, $2\Delta f, \ldots, n\Delta f$, respectively. Each one of these components is an FM signal separated from the others. Thus, a bandpass filter centering at $n\omega_c$ can recover an FM signal whose instantaneous frequency has been multiplied by a factor of n. These devices, consisting of nonlinearity and bandpass filters, are known as **frequency multipliers**. In fact, a frequency multiplier can increase both the carrier frequency and the frequency deviation by an integer n. Thus, if we want a twelfth-fold increase in the frequency deviation, we can use a twelfth-order nonlinear device or two second-order and one third-order devices in cascade. The output has a bandpass filter centered at $12\omega_c$, so that it selects only the appropriate term, whose carrier frequency as well as the frequency deviation Δf are 12 times the original values.

This forms the basis of the Armstrong indirect frequency modulator. First, generate an NBFM approximately. Then multiply the NBFM frequency and limit its amplitude variation. Generally, we require to increase Δf by a very large factor n. This increases the carrier frequency also by n. Such a large increase in the carrier frequency may not be needed. In this case we can apply frequency mixing (see Example 4.2, Fig. 4.7) to shift down the carrier frequency to the desired value.

A simplified diagram of a commercial FM transmitter using Armstrong's method is shown in Fig. 5.10. The final output is required to have a carrier frequency of 91.2 MHz and $\Delta f = 75$ kHz. We begin with NBFM with a carrier frequency $f_{c_1} = 200$ kHz generated by a crystal oscillator. This frequency is chosen because it is easy to construct stable crystal oscillators as well as balanced modulators at this frequency. To maintain $\beta \ll 1$, as required in NBPM, the

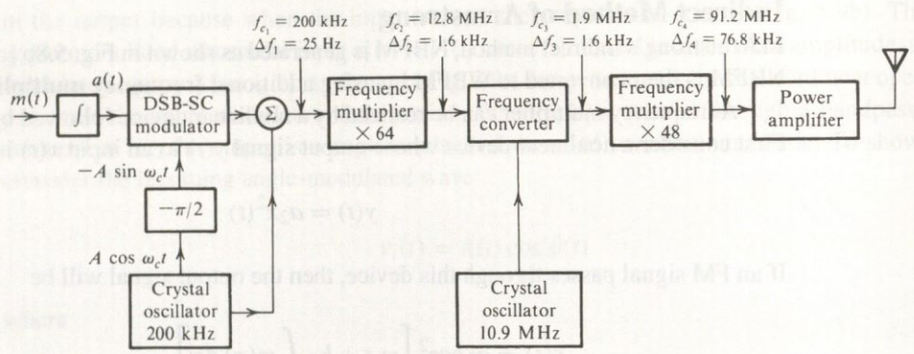

Figure 5.10 Block diagram of the Armstrong indirect FM transmitter.

deviation Δf is chosen to be 25 Hz. For tone modulation, $\beta = \Delta f / f_m$. The baseband spectrum (required for high-fidelity purposes) ranges from 50 Hz to 15 kHz. The choice of $\Delta f = 25$ Hz is reasonable because it gives $\beta = 0.5$ for the worst possible case ($f_m = 50$).

To achieve $\Delta f = 75$ kHz, we need a multiplication of $75{,}000/25 = 3000$. This can be done by two multiplier stages, of 64 and 48, as shown in Fig. 5.10, giving a total multiplication of $64 \times 48 = 3072$, and $\Delta f = 76.8$ kHz.* The multiplication is effected by using frequency doublers and triplers in cascade, as needed. Thus, a multiplication of 64 can be obtained by six doublers in cascade, and a multiplication of 48 can be obtained by four doublers and a tripler in cascade. Multiplication of $f_c = 200$ kHz by 3072, however, would yield a final carrier of about 600 MHz. This problem is solved by using a frequency translation, or conversion, after the first multiplier (Fig. 5.10). The first multiplication by 64 results in the carrier frequency $f_{c_2} = 200\,\text{kHz} \times 64 = 12.8$ MHz, and the carrier deviation $\Delta f_2 = 25 \times 64 = 1.6$ kHz. We now use a frequency converter (or mixer) with carrier frequency 10.9 MHz to shift the entire spectrum. This results in a new carrier frequency $f_{c_3} = 12.8 - 10.9 = 1.9$ MHz. The frequency converter shifts the entire spectrum without altering Δf. Hence, $\Delta f_3 = 1.6$ kHz. Further multiplication, by 48, yields $f_{c_4} = 1.9 \times 48 = 91.2$ MHz and $\Delta f_4 = 1.6 \times 48 = 76.8$ kHz.

This scheme has an advantage of frequency stability, but it suffers from inherent noise caused by excessive multiplication and distortion at lower modulating frequencies, where $\Delta f / f_m$ is not small enough.

Example 5.6 Discuss the nature of distortion inherent in the Armstrong indirect FM generator.

Two kinds of distortion arise in this scheme: amplitude distortion and frequency distortion. The NBFM wave is given by [Eq. (5.10)]

$$\varphi_{\text{FM}}(t) = A[\cos \omega_c t - k_f a(t) \sin \omega_c t]$$
$$= AE(t) \cos [\omega_c t + \theta(t)]$$

where

$$E(t) = \sqrt{1 + k_f^2 a^2(t)} \qquad \text{and} \qquad \theta(t) = \tan^{-1}[k_f a(t)]$$

* If we wish Δf to be exactly 75 kHz instead of 76.8 kHz, we must reduce the narrowband Δf from 25 Hz to $25(75/76.8) = 24.41$ Hz.

Amplitude distortion occurs because the amplitude $AE(t)$ of the modulated waveform is not constant. This is not a serious problem because amplitude variations can be eliminated by a bandpass limiter, as discussed earlier in the section (see also Fig. 5.9). Ideally, $\theta(t)$ should be $k_f a(t)$. Instead, the phase $\theta(t)$ in the preceding equation is

$$\theta(t) = \tan^{-1}[k_f a(t)]$$

and the instantaneous frequency $\omega_i(t)$ is

$$\begin{aligned}\omega_i(t) = \dot{\theta}(t) &= \frac{k_f \dot{a}(t)}{1 + k_f^2 a^2(t)} \\ &= \frac{k_f m(t)}{1 + k_f^2 a^2(t)} \\ &= k_f m(t)[1 - k_f^2 a^2(t) + k_f^4 a^4(t) - \cdots]\end{aligned}$$

Ideally, the instantaneous frequency should be $k_f m(t)$. The remaining terms in this equation are the distortion.

Let us investigate the effect of this distortion in tone modulation where $m(t) = \alpha \cos \omega_m t$, $a(t) = \alpha \sin \omega_m t/\omega_m$, and the modulation index $\beta = \alpha k_f/\omega_m$:

$$\omega_i(t) = \beta\omega_m \cos \omega_m t(1 - \beta^2 \sin^2 \omega_m t + \beta^4 \sin^4 \omega_m t - \cdots)$$

It is evident from this equation that the scheme has odd-harmonic distortion, the most important term being the third harmonic. Ignoring the remaining terms, this equation becomes

$$\begin{aligned}\omega_i(t) &\simeq \beta\omega_m \cos \omega_m t(1 - \beta^2 \sin^2 \omega_m t) \\ &= \underbrace{\beta\omega_m \left(1 - \frac{\beta^2}{4}\right) \cos \omega_m t}_{\text{desired}} + \underbrace{\frac{\beta^3 \omega_m}{4} \cos 3\omega_m t}_{\text{distortion}}\end{aligned}$$

The ratio of the third-harmonic distortion to the desired signal can be found for the generator in Fig. 5.10. For the NBFM stage,

$$\beta B = \Delta f_1 = 25 \text{ Hz}$$

Hence, the worst possible case occurs at the lower modulation frequency. For example, if the tone frequency is only 50 Hz, then $\beta = 0.5$. In this case the third-harmonic distortion is 1/15, or 6.67%.

Direct Generation

In a voltage-controlled oscillator (VCO), the frequency is controlled by an external voltage. The oscillation frequency varies linearly with the control voltage. We can generate an FM wave by using the modulating signal $m(t)$ as a control signal. This gives

$$\omega_i(t) = \omega_c + k_f m(t)$$

One can construct a VCO using an operational amplifier and a hysteretic comparator,[9] (such as a Schmitt trigger circuit). Another way of accomplishing the same goal is to vary one of the reactive parameters (C or L) of the resonant circuit of an oscillator. A *reverse-biased semiconductor diode* acts as a capacitor whose capacitance varies with the bias voltage. The capacitance of these diodes, known under several trade names (e.g., Varicap, Varactor, Voltacap), can be approximated as a linear function of the bias voltage $m(t)$ over a limited range. In Hartley or Colpitt oscillators, for instance, the frequency of oscillation is given by

$$\omega_0 = \frac{1}{\sqrt{LC}}$$

If the capacitance C is varied by the modulating signal $m(t)$, that is, if

$$C = C_0 - km(t)$$

then

$$\begin{aligned}\omega_0 &= \frac{1}{\sqrt{LC_0\left[1 - \dfrac{km(t)}{C_0}\right]}} \\ &= \frac{1}{\sqrt{LC_0}\left[1 - \dfrac{km(t)}{C_0}\right]^{1/2}} \\ &\approx \frac{1}{\sqrt{LC_0}}\left[1 + \frac{km(t)}{2C_0}\right] \qquad \frac{km(t)}{C_0} \ll 1\end{aligned}$$

Here we have applied the binomial approximation

$$(1+x)^n \approx 1 + nx \qquad |x| \ll 1$$

with $n = 1/2$. Thus,

$$\begin{aligned}\omega_0 &= \omega_c\left[1 + \frac{km(t)}{2C_0}\right] \quad \text{where} \quad \omega_c = \frac{1}{\sqrt{LC_0}} \\ &= \omega_c + k_f m(t) \qquad \text{with} \quad k_f = \frac{k\omega_c}{2C_0}\end{aligned}$$

Because $C = C_0 - km(t)$, the maximum capacitance deviation is

$$\Delta C = km_p = \frac{2k_f C_0 m_p}{\omega_c}$$

Hence,

$$\frac{\Delta C}{C_0} = \frac{2k_f m_p}{\omega_c} = \frac{2\Delta f}{f_c}$$

In practice, $\Delta f / f_c$ is usually small, and, hence, ΔC is a small fraction of C_0, which helps limit the harmonic distortion that arises because of the approximation used in this derivation.

Example 5.7

Figure 5.11 Designing an Armstrong indirect modulator.

We may also generate direct FM by using a saturable core reactor, where the inductance of a coil is varied by a current through a second coil (also wound around the same core). This results in a variable inductor whose inductance is proportional to the current in the second coil.

Direct FM generation generally produces sufficient frequency deviation and requires little frequency multiplication. But this method has poor frequency stability. In practice, feedback is used to stabilize the frequency. The output frequency is compared with a constant frequency generated by a stable crystal oscillator. An error signal (error in frequency) is detected and fed back to the oscillator to correct the error.

Features of Angle Modulation

FM (like angle modulation in general) has a number of unique features that recommend it for various radio systems. The transmission bandwidth of AM systems cannot be changed. Because of this, AM systems do not have the feature of exchanging signal power for transmission bandwidth. Pulse-coded modulation (PCM) systems (Chapter 6) have such a feature, and so do angle-modulated systems. In angle modulation, the transmission bandwidth can be adjusted by adjusting Δf. It is shown in Chapter 10 that for angle-modulated systems, the SNR is roughly proportional to the square of the transmission bandwidth B_T. Recall that in PCM, the SNR varies exponentially with B_T and is, therefore, superior to angle modulation.

Example 5.7 Design an Armstrong indirect FM modulator to generate an FM signal with carrier frequency 97.3 MHz and $\Delta f = 10.24$ kHz. A NBFM generator of $f_{c_1} = 20$ kHz and $\Delta f = 5$ Hz is available. Only frequency doublers can be used as multipliers. Additionally, a local oscillator (LO) with adjustable frequency between 400 and 500 kHz is readily available for frequency mixing.

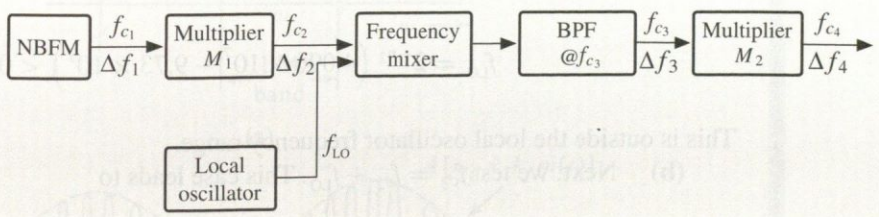

Figure 5.11 Designing an Armstrong indirect modulator.

The modulator is shown in Fig. 5.11. We need to determine M_1, M_2, and f_{LO}. First, the NBFM generator generates

$$f_{c_1} = 20,000 \qquad \text{and} \qquad \Delta f_1 = 5$$

The final WBFM should have

$$f_{c_4} = 97.3 \times 10^6 \qquad \Delta f_4 = 10,240$$

We first find the total factor of frequency multiplication needed as

$$M_1 \cdot M_2 = \frac{\Delta f_4}{\Delta f_1} = 2048 = 2^{11} \tag{5.25}$$

Because only frequency doublers can be used, we have three equations:

$$M_1 = 2^{n_1}$$
$$M_2 = 2^{n_2}$$
$$n_1 + n_2 = 11$$

It is also clear that

$$f_{c_2} = 2^{n_1} f_{c_1} \quad \text{and} \quad f_{c_4} = 2^{n_2} f_{c_3}$$

To find f_{LO}, there are three possible relationships:

$$f_{c_3} = f_{c_2} \pm f_{LO} \quad \text{and} \quad f_{c_3} = f_{LO} - f_{c_2}$$

Each should be tested to determine the one that will fall in

$$400,000 \leq f_{LO} \leq 500,000$$

(a) First, we test $f_{c_3} = f_{c_2} - f_{LO}$. This case leads to

$$\begin{aligned} 97.3 \times 10^6 &= 2^{n_2}\left(2^{n_1} f_{c_1} - f_{LO}\right) \\ &= 2^{n_1+n_2} f_{c_1} - 2^{n_2} f_{LO} \\ &= 2^{11} 20 \times 10^3 - 2^{n_2} f_{LO} \end{aligned}$$

Thus, we have

$$f_{LO} = 2^{-n_2}\left(4.096 \times 10^7 - 9.73 \times 10^7\right) < 0$$

This is outside the local oscillator frequency range.

(b) Next, we test $f_{c_3} = f_{c_2} + f_{LO}$. This case leads to

$$\begin{aligned} 97.3 \times 10^6 &= 2^{n_2}\left(2^{n_1} f_{c_1} + f_{LO}\right) \\ &= 2^{11} 20 \times 10^3 + 2^{n_2} f_{LO} \end{aligned}$$

Thus, we have

$$f_{LO} = 2^{-n_2}\left(5.634 \times 10^7\right)$$

If $n_2 = 7$, then $f_{LO} = 440$ kHz, which is within the realizable range of the local oscillator.

(c) If we choose $f_{c_3} = f_{LO} - f_{c_2}$, then we have

$$\begin{aligned} 97.3 \times 10^6 &= f_{LO} - 2^{n_2} 2^{n_1} f_{c_1} \\ &= 2^{n_2} f_{LO} - 2^{11}(20 \times 10^3) \end{aligned}$$

Thus, we have

$$f_{LO} = 2^{-n_2}\left(13.826 \times 10^7\right)$$

No integer n_2 will lead to a realizable f_{LO}.

Thus, the final design is $M_1 = 16$, $M_2 = 128$, and $f_{LO} = 440$ kHz.

5.4 DEMODULATION OF FM SIGNALS

The information in an FM signal resides in the instantaneous frequency $\omega_i = \omega_c + k_f m(t)$. Hence, a frequency-selective network with a transfer function of the form $|H(f)| = 2a\pi f + b$ over the FM band would yield an output proportional to the instantaneous frequency (Fig. 5.12a).* There are several possible circuits with such characteristics. The simplest among them is an ideal differentiator with the transfer function $j2\pi f$.

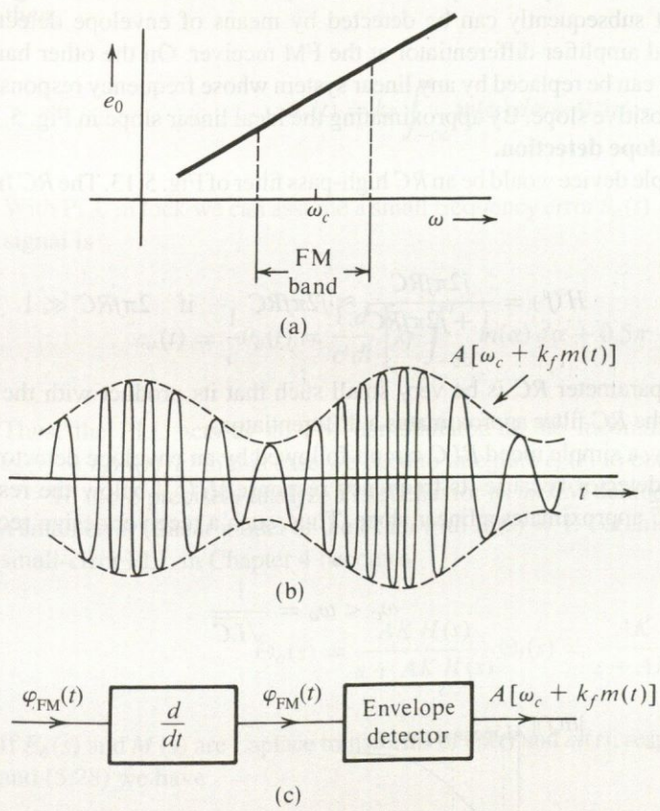

Figure 5.12 (a) FM demodulator frequency response. (b) Output of a differentiator to the input FM wave. (c) FM demodulation by direct differentiation.

* Provided the variations of ω_i are slow in comparison to the time constant of the network.

If we apply $\varphi_{\text{FM}}(t)$ to an ideal differentiator, the output is

$$\dot{\varphi}_{\text{FM}}(t) = \frac{d}{dt}\left\{A\cos\left[\omega_c t + k_f\int_{-\infty}^{t} m(\alpha)\,d\alpha\right]\right\}$$

$$= A\left[\omega_c + k_f m(t)\right]\sin\left[\omega_c t + k_f\int_{-\infty}^{t} m(\alpha)\,d(\alpha) - \pi\right] \tag{5.26}$$

Both the amplitude and the frequency of the signal $\dot{\varphi}_{\text{FM}}(t)$ are modulated (Fig. 5.12b), the envelope being $A[\omega_c + k_f m(t)]$. Because $\Delta\omega = k_f m_p < \omega_c$, we have $\omega_c + k_f m(t) > 0$ for all t, and $m(t)$ can be obtained by envelope detection of $\dot{\varphi}_{\text{FM}}(t)$ (Fig. 5.12c).

The amplitude A of the incoming FM carrier must be constant. If the amplitude A were not constant, but a function of time, there would be an additional term containing dA/dt on the right-hand side of Eq. (5.26). Even if this term were neglected, the envelope of $\dot{\varphi}_{\text{FM}}(t)$ would be $A(t)[\omega_c + k_f m(t)]$, and the envelope-detector output would be proportional to $m(t)A(t)$, still leading to distortions. Hence, it is essential to maintain A constant. Several factors, such as channel noise and fading, cause A to vary. This variation in A should be suppressed via the bandpass limiter (discussed earlier in Sec. 5.3) before the signal is applied to the FM detector.

Practical Frequency Demodulators

The differentiator is only one way to convert frequency variation of FM signals into amplitude variation that subsequently can be detected by means of envelope detectors. One can use an operational amplifier differentiator at the FM receiver. On the other hand, the role of the differentiator can be replaced by any linear system whose frequency response contains a linear segment of positive slope. By approximating the ideal linear slope in Fig. 5.12(a), this method is known as **slope detection.**

One simple device would be an RC high-pass filter of Fig. 5.13. The RC frequency response is simply

$$H(f) = \frac{j2\pi fRC}{1 + j2\pi fRC} \approx j2\pi fRC \qquad \text{if} \quad 2\pi fRC \ll 1$$

Thus, if the parameter RC is be very small such that its product with the carrier frequency $\omega_c RC \ll 1$, the RC filter approximates a differentiator.

Similarly, a simple tuned RLC circuit followed by an envelope detector can also serve as a frequency detector because its frequency response $|H(f)|$ below the resonance frequency $\omega_o = 1/\sqrt{LC}$ approximates a linear slope. Thus, such a receiver design requires that

$$\omega_c < \omega_o = \frac{1}{\sqrt{LC}}$$

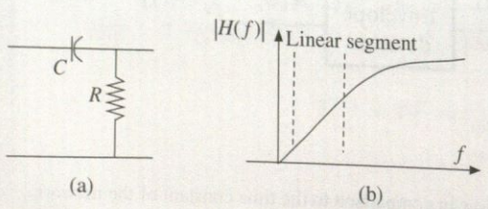

Figure 5.13
(a) RC high-pass filter.
(b) Segment of positive slope in amplitude response.

Because the operation is on the slope of $|H(f)|$, this method is also called **slope detection**. Since, however, the slope of $|H(f)|$ is linear over only a small band, there is considerable distortion in the output. This fault can be partially corrected by a **balanced discriminator** formed by two slope detectors. Another balanced demodulator, the **ratio detector**, also widely used in the past, offers better protection against carrier amplitude variations than does the discriminator. For many years ratio detectors were standard in almost all FM receivers.[10]

Zero-crossing detectors are also used because of advances in digital integrated circuits. The first step is to use the amplitude limiter of Fig. 5.9(a) to generate the rectangular pulse output of Fig. 5.9(c). The resulting rectangular pulse train of varying width can then be applied to trigger a digital counter. These are the **frequency counters** designed to measure the instantaneous frequency from the number of zero crossings. The rate of zero crossings is equal to the instantaneous frequency of the input signal.

FM Demodulation via PLL

Consider a PLL that is in lock with input signal $\sin[\omega_c t + \theta_i(t)]$ and output error signal $e_o(t)$. When the input signal is an FM signal,

$$\theta_i(t) = k_f \int_{-\infty}^{t} m(\alpha)\, d\alpha + \frac{\pi}{2} \tag{5.27}$$

then,

$$\theta_o(t) = k_f \int_{-\infty}^{t} m(\alpha)\, d\alpha + 0.5\pi - \theta_e(t)$$

With PLL in lock we can assume a small frequency error $\dot{\theta}_e(t) \approx 0$. Thus, the loop filter output signal is

$$e_o(t) = \frac{1}{c}\dot{\theta}_o(t) = \frac{1}{c}\frac{d}{dt}\left[k_f \int_{-\infty}^{t} m(\alpha)\, d\alpha + 0.5\pi - \theta_e(t)\right] \simeq \frac{k_f}{c} m(t) \tag{5.28}$$

Thus, the PLL acts as an FM demodulator. If the incoming signal is a PM wave, then $e_o(t) = k_p \dot{m}(t)/c$. In this case we need to integrate $e_o(t)$ to obtain the desired signal $m(t)$.

To more precisely analyze PLL behavior as an FM demodulator, we consider the case of a small error (linear model of the PLL) with $H(s) = 1$. For this case, feedback analysis of the small-error PLL in Chapter 4 becomes

$$\Theta_o(s) = \frac{AK\,H(s)}{s + AK\,H(s)}\Theta_i(s) = \frac{AK}{s + AK}\Theta_i(s)$$

If $E_o(s)$ and $M(s)$ are Laplace transforms of $e_o(t)$ and $m(t)$, respectively, then from Eqs. (5.27) and (5.28) we have

$$\Theta_i(s) = \frac{k_f M(s)}{s} \qquad \text{and} \qquad s\Theta_o(s) = cE_o(s)$$

Hence,

$$E_o(s) = \left(\frac{k_f}{c}\right)\frac{AK}{s+AK}M(s)$$

Thus, the PLL output $e_o(t)$ is a distorted version of $m(t)$ and is equivalent to the output of a single-pole circuit (such as a simple RC circuit) with transfer function $k_f AK/c(s+AK)$ to which $m(t)$ as the input. To reduce distortion, we must choose AK well above the bandwidth of $m(t)$, so that $e_o(t) \simeq k_f m(t)/c$.

In the presence of small noise, the behavior of the PLL is comparable to that of a frequency discriminator. The advantage of the PLL over a frequency discriminator appears only when the noise is large.

5.5 EFFECTS OF NONLINEAR DISTORTION AND INTERFERENCE

Immunity of Angle Modulation to Nonlinearities

A very useful feature of angle modulation is its constant amplitude, which makes it less susceptible to nonlinearities. Consider, for instance, an amplifier with second-order nonlinear distortion whose input $x(t)$ and output $y(t)$ are related by

$$y(t) = a_0 + a_1 x(t) + a_2 x^2(t) + \cdots + a_n x^n(t)$$

Clearly, the first term is the desired signal amplification term, while the remaining terms are the unwanted nonlinear distortion. For the angle modulated signal

$$x(t) = A\cos[\omega_c t + \psi(t)]$$

trigonometric identities can be applied to rewrite the nonideal system output $y(t)$ as

$$\begin{aligned} y(t) = c_o + c_1\cos[\omega_c t + \psi(t)] + c_2\cos[2\omega_c t + 2\psi(t)] \\ + \cdots + c_n\cos[n\omega_c t + n\psi(t)] \end{aligned}$$

Because sufficiently large ω_c makes each component of $y(t)$ separable in frequency domain, a bandpass filter centered at ω_c with bandwidth equaling to B_{FM} (or B_{PM}) can extract the desired FM signal component $c_1\cos[\omega_c t + \psi(t)]$ without any distortion. This shows that angle-modulated signals are immune to nonlinear distortions.

A similar nonlinearity in AM not only causes unwanted modulation with carrier frequencies $n\omega_c$ but also causes distortion of the desired signal. For instance, if a DSB-SC signal $m(t)\cos\omega_c t$ passes through a nonlinearity $y(t) = a\,x(t) + b\,x^3(t)$, the output is

$$\begin{aligned} y(t) &= a\,m(t)\cos\omega_c t + b\,m^3(t)\cos^3\omega_c t \\ &= \left[a\,m(t) + \frac{3b}{4}m^3(t)\right]\cos\omega_c t + \frac{b}{4}m^3(t)\cos 3\omega_c t \end{aligned}$$

Passing this signal through a bandpass filter still yields $[a\,m(t)+(3b/4)m^3(t)]\cos\omega_c t$. Observe the distortion component $(3b/4)m^3(t)$ present along with the desired signal $a\,m(t)$.

Immunity from nonlinearity is the primary reason for the use of angle modulation in microwave radio relay systems, where power levels are high. This requires highly efficient nonlinear class C amplifiers. In addition, the constant amplitude of FM gives it a kind of immunity to rapid fading. The effect of amplitude variations caused by rapid fading can be eliminated by using automatic gain control and bandpass limiting. These advantages made FM attractive as the technology behind the first-generation (1G) cellular phone system.

The same advantages of FM also make it attractive for microwave radio relay systems. In the legacy analog long-haul telephone systems, several channels are multiplexed by means of SSB signals to form L-carrier signals. The multiplexed signals are frequency-modulated and transmitted over a microwave radio relay system with many links in tandem. In this application, however, FM is used not to reduce noise effects but to realize other advantages of constant amplitude, and, hence, NBFM rather than WBFM is used.

Interference Effect

Angle modulation is also less vulnerable than AM to small-signal interference from adjacent channels.

Let us consider the simple case of the interference of an unmodulated carrier $A\cos\omega_c t$ with another sinusoid $I\cos(\omega_c+\omega)t$. The received signal $r(t)$ is

$$\begin{aligned} r(t) &= A\cos\omega_c t + I\cos(\omega_c+\omega)t \\ &= (A+I\cos\omega t)\cos\omega_c t - I\sin\omega t\sin\omega_c t \\ &= E_r(t)\cos[\omega_c t+\psi_d(t)] \end{aligned}$$

where

$$\psi_d(t) = \tan^{-1}\frac{I\sin\omega t}{A+I\cos\omega t}$$

When the interfering signal is small in comparison to the carrier ($I \ll A$),

$$\psi_d(t) \simeq \frac{I}{A}\sin\omega t \tag{5.29}$$

The phase of $E_r(t)\cos[\omega_c t+\psi_d(t)]$ is $\psi_d(t)$, and its instantaneous frequency is $\omega_c+\dot{\psi}_d(t)$. If the signal $E_r(t)\cos[\omega_c t+\psi_d(t)]$ is applied to an ideal phase demodulator, the output $y_d(t)$ would be $\psi_d(t)$. Similarly, the output $y_d(t)$ of an ideal frequency demodulator would be $\dot{\psi}_d(t)$. Hence,

$$y_d(t) = \frac{I}{A}\sin\omega t \qquad \text{for PM} \tag{5.30}$$

$$y_d(t) = \frac{I\omega}{A}\cos\omega t \qquad \text{for FM} \tag{5.31}$$

Observe that in either case, the interference output is inversely proportional to the carrier amplitude A. Thus, the larger the carrier amplitude A, the smaller the interference effect. This behavior is very different from that in AM signals, where the interference output is independent

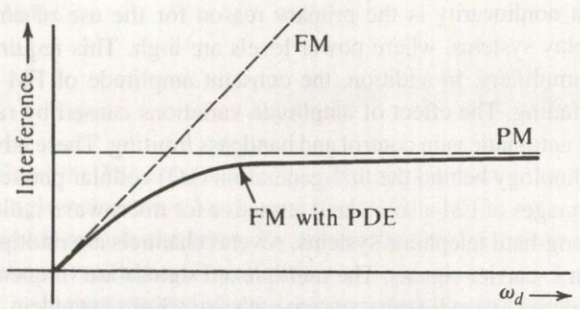

Figure 5.14 Effect of interference in PM, FM, and FM with preemphasis-deemphasis (PDE).

of the carrier amplitude.* Hence, angle-modulated systems are much better than AM systems at suppressing weak interference ($I \ll A$).

Because of the suppression of weak interference in FM, we observe what is known as the **capture effect** when listening to FM radios. For two transmitters with carrier frequency separation less than the audio range, instead of getting interference, we observe that the stronger carrier effectively suppresses (captures) the weaker carrier. Subjective tests show that an interference level as low as 35 dB in the audio signals can cause objectionable effects. Hence, in AM, the interference level should be kept below 35 dB. On the other hand, for FM, because of the capture effect, the interference level need only be below 6 dB.

The interference amplitude (I/A for PM and $I\omega/A$ for FM) vs. ω at the receiver output is shown in Fig. 5.14. The interference amplitude is constant for all ω in PM but increases linearly with ω in FM.†

Interference due to Channel Noise

The channel noise acts as interference in an angle-modulated signal. We shall consider the most common form of noise, white noise, which has a constant power spectral density. Such a noise may be considered as a sum of sinusoids of all frequencies in the band. All components have the same amplitudes (because of uniform density). This means I is constant for all ω, and the amplitude spectrum of the interference at the receiver output is as shown in Fig. 5.14. The interference amplitude spectrum is constant for PM, and increases linearly with ω for FM.

Preemphasis and Deemphasis in FM Broadcasting

Figure 5.14 shows that in FM, the interference (the noise) increases linearly with frequency, and the noise power in the receiver output is concentrated at higher frequencies. A glance at Fig. 4.18b shows that the PSD of an audio signal $m(t)$ is concentrated at lower frequencies below 2.1 kHz. Thus, the noise PSD is concentrated at higher frequencies, where $m(t)$ is

Figure 5.15 Preemphasis-deemphasis in an FM system.

Figure 5.16 (a) Preemphasis filter and (b) its frequency response. (c) Deemphasis filter and (d) its frequency response.

---

* For instance, an AM signal with an interfering sinusoid $I\cos(\omega_c + \omega)t$ is given by

$$r(t) = [A + m(t)]\cos\omega_c t + I\cos(\omega_c + \omega)t$$
$$= [A + m(t) + I\cos\omega t]\cos\omega_c t - I\sin\omega t\sin\omega_c t$$

The envelope of this signal is

$$E(t) = \{[A + m(t) + I\cos\omega t]^2 + I^2\sin^2\omega t\}^{1/2} \approx A + m(t) + I\cos\omega t \qquad I \ll A$$

Thus the interference signal at the envelope detector output is $I\cos\omega t$, which is independent of the carrier amplitude A. We obtain the same result when synchronous demodulation is used. We come to a similar conclusion for AM-SC systems.

† The results in Eqs. (5.30) and (5.31) can be readily extended to more than one interfering sinusoid. The system behaves linearly for multiple interfering sinusoids provided their amplitudes are very small in comparison to the carrier amplitude.

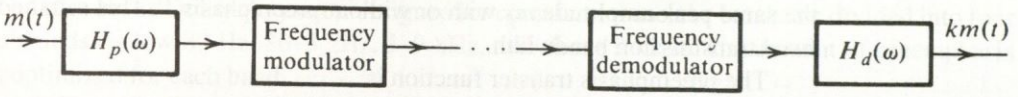

Figure 5.15 Preemphasis-deemphasis in an FM system.

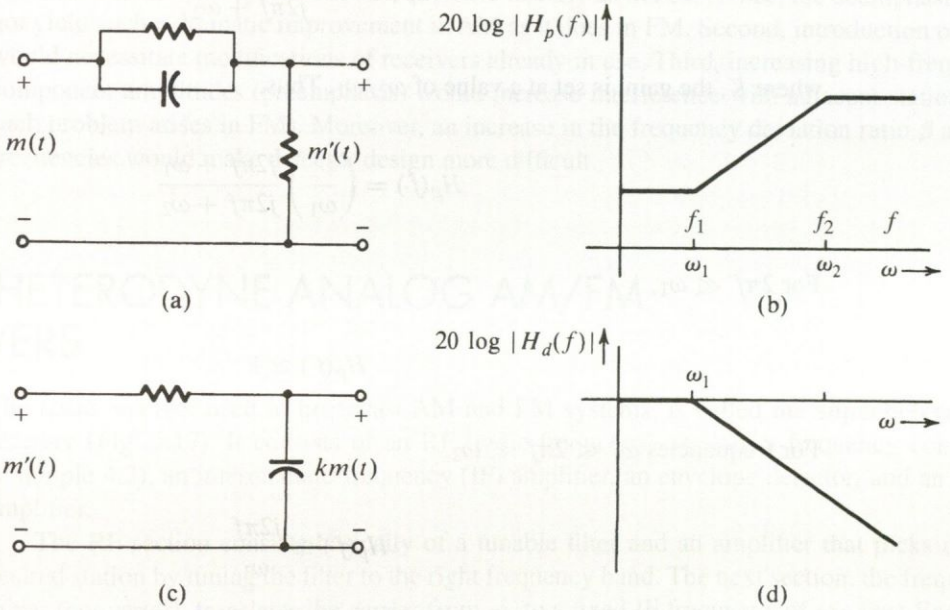

Figure 5.16 (a) Preemphasis filter and (b) its frequency response. (c) Deemphasis filter and (d) its frequency response.

the weakest. This may seem like a disaster. But actually, in this very situation there is a hidden opportunity to reduce noise greatly. The process, shown in Fig. 5.15, works as follows. At the transmitter, the weaker high-frequency components (beyond 2.1 kHz) of the audio signal $m(t)$ are boosted before modulation by a **preemphasis** filter of transfer function $H_p(f)$. At the receiver, the demodulator output is passed through a **deemphasis** filter of transfer function $H_d(f) = 1/H_p(f)$. Thus, the deemphasis filter undoes the preemphasis by attenuating (deemphasizing) the higher frequency components (beyond 2.1 kHz), and thereby restores the original signal $m(t)$. The noise, however, enters at the channel, and therefore has not been preemphasized (boosted). However, it passes through the deemphasis filter, which attenuates its higher frequency components, where most of the noise power is concentrated (Fig. 5.14). Thus, the process of preemphasis-deemphasis (PDE) leaves the desired signal untouched but reduces the noise power considerably.

Preemphasis and Deemphasis Filters

Figure 5.14 provides an opportunity to preemphasis. The FM has smaller interference than PM at lower frequencies, while the opposite is true at higher frequencies. If we can make our system behave like FM at lower frequencies and behave like PM at higher frequencies, we will have the best of both worlds. This is accomplished by a system used in commercial broadcasting (Fig. 5.15) with the preemphasis (before modulation) and deemphasis (after demodulation) filters $H_p(f)$ and $H_d(f)$ shown in Fig. 5.16. The frequency f_1 is 2.1 kHz, and f_2 is typically 30 kHz or more (well beyond audio range), so that f_2 does not even enter into the picture. These filters can be realized by simple RC circuits (Fig. 5.16). The choice of $f_1 = 2.1$ kHz was apparently made on an experimental basis. It was found that this choice of f_1 maintained

the same peak amplitude m_p with or without preemphasis.[11] This satisfied the constraint of a fixed transmission bandwidth.

The preemphasis transfer function is

$$H_p(f) = K\frac{j2\pi f + \omega_1}{j2\pi f + \omega_2} \tag{5.32a}$$

where K, the gain, is set at a value of ω_2/ω_1. Thus,

$$H_p(f) = \left(\frac{\omega_2}{\omega_1}\right)\frac{j2\pi f + \omega_1}{j2\pi f + \omega_2} \tag{5.32b}$$

For $2\pi f \ll \omega_1$,

$$H_p(f) \simeq 1 \tag{5.32c}$$

For frequencies $\omega_1 \ll 2\pi f \ll \omega_2$,

$$H_p(f) \simeq \frac{j2\pi f}{\omega_1} \tag{5.32d}$$

Thus, the preemphasizer acts as a differentiator at intermediate frequencies (2.1–15 kHz), which effectively makes the scheme PM over these frequencies. This means that FM with PDE is FM over the modulating-signal frequency range of 0 to 2.1 kHz and is nearly PM over the range of 2.1 to 15 kHz, as desired.

The deemphasis filter $H_d(f)$ is given by

$$H_d(f) = \frac{\omega_1}{j2\pi f + \omega_1}$$

Note that for $2\pi f \ll \omega_2$, $H_p(f) \simeq (j2\pi f + \omega_1)/\omega_1$. Hence, $H_p(f)H_d(f) \simeq 1$ over the baseband of 0 to 15 kHz.

Optimum PDE filters are discussed in Chapter 10. For historical and practical reasons, optimum PDE filters are not used in practice. It can be shown that the PDE enhances the SNR by 13.27 dB (a power ratio of 21.25).

The side benefit of PDE is improvement in the interference characteristics. Because the interference (from unwanted signals and the neighboring stations) enters after the transmitter stage, it undergoes only the deemphasis operation, not the boosting, or preemphasis. Hence, the interference amplitudes for frequencies beyond 2.1 kHz undergo attenuation that is roughly linear with frequency.

The PDE method of noise reduction is not limited to FM broadcast. It is also used in audiotape recording and in (analog) phonograph recording, where the hissing noise is also concentrated at the high-frequency end. A sharp, hissing sound is caused by irregularities in the recording material. The **Dolby noise reduction** systems for audiotapes operates on the same principle, although the Dolby-A system is somewhat more elaborate. In the Dolby-B and Dolby-C systems, the band is divided into two subbands (below and above 3 kHz instead of

5.6 SUPE RECE

Figure 5.17 Superheterodyne receiver.

2.1 kHz). In the Dolby-A system, designed for commercial use, the bands are divided into four subbands (below 80 Hz, 80–3 kHz, 3–9 kHz, and above 9 kHz). The amount of preemphasis is optimized for each band.

We could also use PDE in AM broadcasting to improve the output SNR. In practice, however, this is not done for several reasons. First, the output noise amplitude in AM is constant with frequency and does not increase linearly as in FM. Hence, the deemphasis does not yield such a dramatic improvement in AM as it does in FM. Second, introduction of PDE would necessitate modifications of receivers already in use. Third, increasing high-frequency component amplitudes (preemphasis) would increase interference with adjacent stations (no such problem arises in FM). Moreover, an increase in the frequency deviation ratio β at high frequencies would make detector design more difficult.

5.6 SUPERHETERODYNE ANALOG AM/FM RECEIVERS

The radio receiver used in broadcast AM and FM systems, is called the **superheterodyne** receiver (Fig. 5.17). It consists of an RF (radio-frequency) section, a frequency converter (Example 4.2), an intermediate-frequency (IF) amplifier, an envelope detector, and an audio amplifier.

The RF section consists basically of a tunable filter and an amplifier that picks up the desired station by tuning the filter to the right frequency band. The next section, the frequency mixer (converter), translates the carrier from ω_c to a fixed IF frequency of ω_{IF} (see Example 4.2 for frequency conversion). For this purpose, the receiver uses a local oscillator whose frequency f_{LO} is exactly f_{IF} above the incoming carrier frequency f_c; that is,

$$f_{LO} = f_c + f_{IF}$$

The simultaneous tuning of the local oscillator and the RF tunable filter is done by one joint knob. Tuning capacitors in both circuits are ganged together and are designed so that the tuning

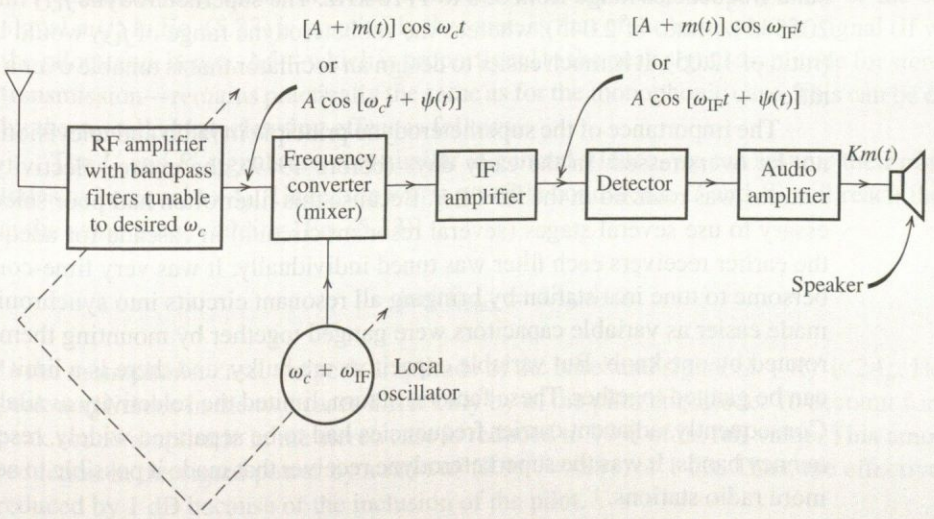

Figure 5.17 Superheterodyne receiver.

frequency of the local oscillator is always f_{IF} Hz above the tuning frequency f_c of the RF filter. This means every station that is tuned in is translated to a fixed carrier frequency of f_{IF} Hz by the frequency converter for subsequent processing at IF.

This superheterodyne receiver structure is broadly utilized in most broadcast systems. The intermediate frequencies are chosen to be 455 kHz (AM radio), 10.7 MHz (FM radio), and 38 MHz (TV reception).

As discovered by Armstrong for AM signals, the translation of all stations to a fixed intermediate frequency ($f_{IF} = 455$ kHz for AM) allows us to obtain adequate selectivity. It is difficult to design precise bandpass filters of bandwidth 10 kHz (the modulated audio spectrum) if the center frequency f_c is very high. This is particularly true in the case of tunable filters. Hence, the RF filter cannot provide adequate selectivity against adjacent channels. But when this signal is translated to an IF frequency by a converter, it is further amplified by an IF amplifier (usually a three-stage amplifier), which does have good selectivity. This is because the IF frequency is reasonably low; moreover, its center frequency is fixed and factory-tuned. Hence, the IF section can effectively suppress adjacent-channel interference because of its high selectivity. It also amplifies the signal for envelope detection.

In reality, the entire selectivity is practically realized in the IF section; the RF section plays a negligible role. The main function of the RF section is image frequency suppression. As observed in Example 4.2, the output of the mixer, or converter, consists of components of the difference between the incoming (f_c) and the local oscillator frequencies (f_{LO}) (i.e., $f_{IF} = |f_{LO} - f_c|$). Now, consider the AM example. If the incoming carrier frequency $f_c = 1000$ kHz, then $f_{LO} = f_c + f_{RF} = 1000 + 455 = 1455$ kHz. But another carrier, with $f_c' = 1455 + 455 = 1910$ kHz, will also be picked up because the difference $f_c' - f_{LO}$ is also 455 kHz. The station at 1910 kHz is said to be the **image** of the station of 1000 kHz. AM stations that are $2f_{IF} = 910$ kHz apart are called **image stations** and both would appear simultaneously at the IF output, were it not for the RF filter at receiver input. The RF filter may provide poor selectivity against adjacent stations separated by 10 kHz, but it can provide reasonable selectivity against a station separated by 910 kHz. Thus, when we wish to tune in a station at 1000 kHz, the RF filter, tuned to 1000 kHz, provides adequate suppression of the image station at 1910 kHz.

The receiver (Fig. 5.17) converts the incoming carrier frequency to the IF by using a local oscillator of frequency f_{LO} higher than the incoming carrier frequency and, hence, is called a superheterodyne receiver. We pick f_{LO} higher than f_c because this leads to a smaller tuning ratio of the maximum to minimum tuning frequency for the local oscillator. The AM broadcast-band frequencies range from 530 to 1710 kHz. The superheterodyne f_{LO} ranges from 1005 to 2055 kHz (ratio of 2.045), whereas the subheterodyne range of f_{LO} would be 95 to 1145 kHz (ratio of 12.05). It is much easier to design an oscillator that is tunable over a smaller frequency ratio.

The importance of the superheterodyne principle in radio and television broadcasting cannot be overstressed. In the early days (before 1919), the entire selectivity against adjacent stations was realized in the RF filter. Because this filter often had poor selectivity, it was necessary to use several stages (several resonant circuits) in cascade for adequate selectivity. In the earlier receivers each filter was tuned individually. It was very time-consuming and cumbersome to tune in a station by bringing all resonant circuits into synchronism. This task was made easier as variable capacitors were ganged together by mounting them on the same shaft rotated by one knob. But variable capacitors are bulky, and there is a limit to the number that can be ganged together. These factors, in turn, limited the selectivity available from receivers. Consequently, adjacent carrier frequencies had to be separated widely, resulting in fewer frequency bands. It was the superheterodyne receiver that made it possible to accommodate many more radio stations.

5.7 FM BROADCASTING SYSTEM

The FCC has assigned a frequency range of 88 to 108 MHz for FM broadcasting, with a separation of 200 kHz between adjacent stations and a peak frequency deviation $\Delta f = 75$ kHz.

A monophonic FM receiver is identical to the superheterodyne AM receiver in Fig. 5.17, except that the intermediate frequency is 10.7 MHz and the envelope detector is replaced by a PLL or a frequency discriminator followed by a deemphasizer.

Earlier FM broadcasts were monophonic. Stereophonic FM broadcasting, in which two audio signals, L (left microphone) and R (right microphone), are used for a more natural effect, was proposed later. The FCC ruled that the stereophonic system had to be compatible with the original monophonic system. This meant that the older monophonic receivers should be able to receive the signal $L + R$, and the total transmission bandwidth for the two signals (L and R) should still be 200 kHz, with $\Delta f = 75$ kHz for the two combined signals. This would ensure that the older receivers could continue to receive monophonic as well as stereophonic broadcasts, although the stereo effect would be absent.

A transmitter and a receiver for a stereo broadcast are shown in Fig. 5.18a and c. At the transmitter, the two signals L and R are added and subtracted to obtain $L + R$ and $L - R$. These signals are preemphasized. The preemphasized signal $(L - R)'$ DSB-SC modulates a carrier of 38 kHz obtained by doubling the frequency of a 19-kHz signal that is used as a pilot. The signal $(L + R)'$ is used directly. All three signals (the third being the pilot) form a composite baseband signal $m(t)$ (Fig. 5.18b),

$$m(t) = (L + R)' + (L - R)' \cos \omega_c t + \alpha \cos \frac{\omega_c t}{2} \tag{5.33}$$

The reason for using a pilot of 19 kHz rather than 38 kHz is that it is easier to separate the pilot at 19 kHz because there are no signal components within 4 kHz of that frequency.

The receiver operation (Fig. 5.18c) is self-explanatory. A monophonic receiver consists of only the upper branch of the stereo receiver and, hence, receives only $L + R$. This is of course the complete audio signal without the stereo effect. Hence, the system is compatible. The pilot is extracted, and (after doubling its frequency) it is used to demodulate coherently the signal $(L - R)' \cos \omega_c t$.

An interesting aspect of stereo transmission is that the peak amplitude of the composite signal $m(t)$ in Eq. (5.33) is practically the same as that of the monophonic signal (if we ignore the pilot), and, hence, Δf—which is proportional to the peak signal amplitude for stereophonic transmission—remains practically the same as for the monophonic case. This can be explained by the so-called **interleaving** effect as follows.

The L' and R' signals are very similar in general. Hence, we can assume their peak amplitudes to be equal to A_p. Under the worst possible conditions, L' and R' will reach their peaks at the same time, yielding [Eq. (5.33)]

$$|m(t)|_{\max} = 2A_p + \alpha$$

In the monophonic case, the peak amplitude of the baseband signal $(L + R)'$ is $2A_p$. Hence, the peak amplitudes in the two cases differ only by α, the pilot amplitude. To account for this, the peak sound amplitude in the stereo case is reduced to 90% of its full value. This amounts to a reduction in the signal power by a ratio of $(0.9)^2 = 0.81$, or 1 dB. Thus, the effective SNR is reduced by 1 dB because of the inclusion of the pilot.

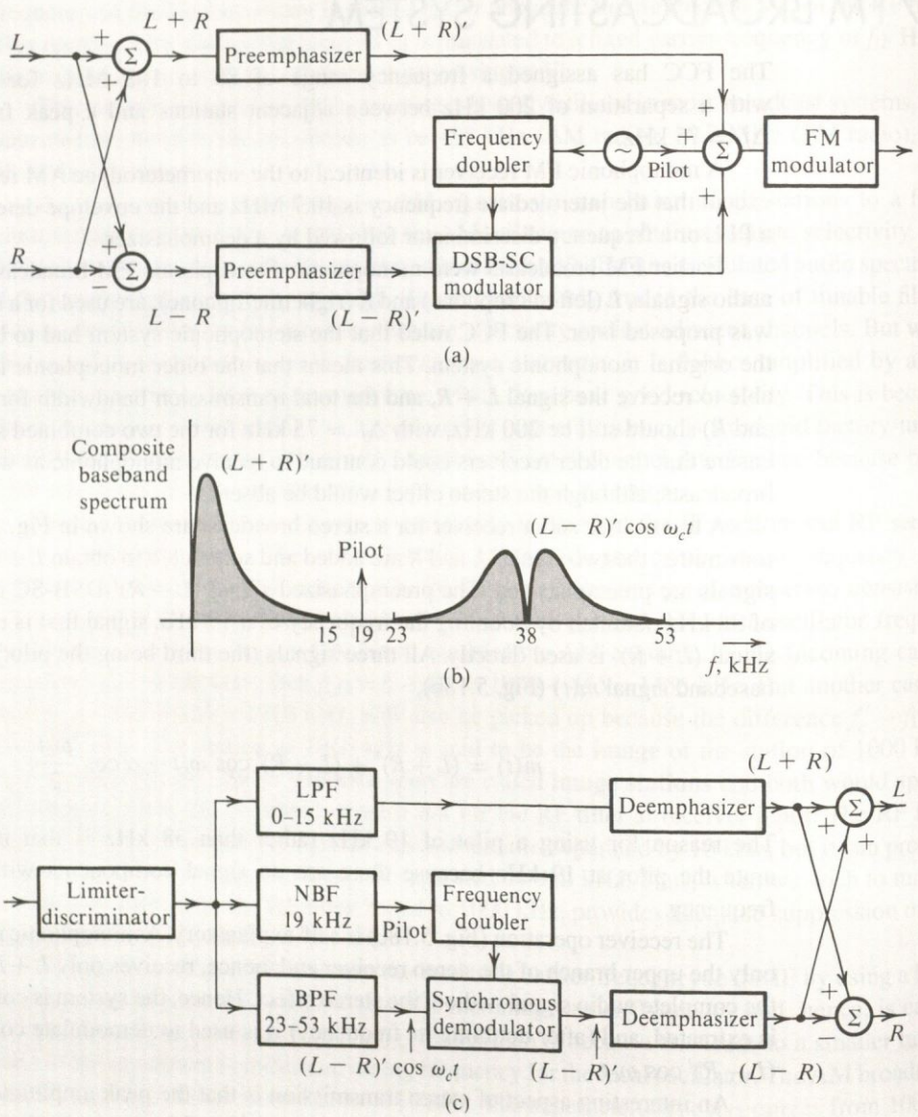

Figure 5.18 (a) FM stereo transmitter. (b) Spectrum of a baseband stereo signal. (c) FM stereo receiver.

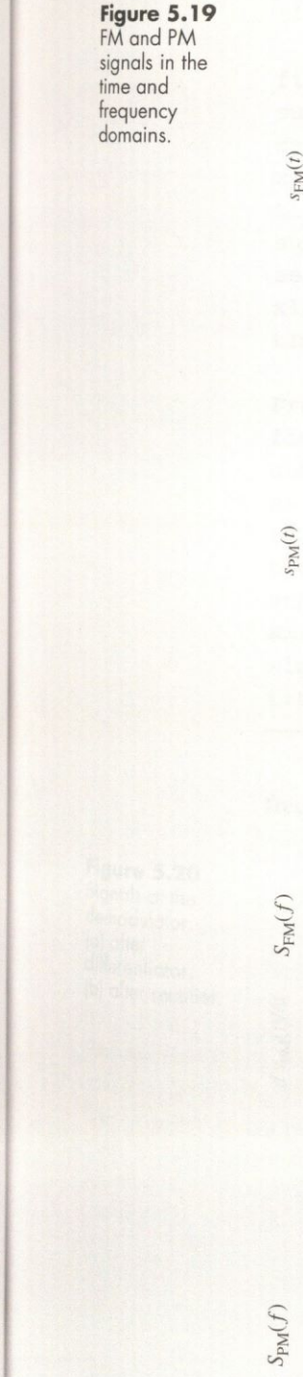

Figure 5.19 FM and PM signals in the time and frequency domains.

5.8 MATLAB EXERCISES

In this section, we use MATLAB to build an FM modulation and demodulation example. The MATLAB program is given by `ExampleFM.m`. Once again use apply the same message signal $m_2(t)$. The FM coefficient is $k_f = 80$ and the PM coefficient is $k_p = \pi$. The carrier frequency remains 300 Hz. The resulting FM and PM signals in the time domain are shown in Fig. 5.19. The corresponding frequency responses are also shown in Fig. 5.19. The frequency domain responses clearly show the much higher bandwidths of the FM and PM signals when compared with amplitude modulations.

```
% (ExampleFM.m)
% This program uses triangl.m to illustrate frequency modulation
% and demodulation
```